UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Département de formation doctorale en informatique

Synthèse Acoustico-Visuelle de la Parole par Séléction d'Unités Bimodales

(Acoustic-Visual Speech Synthesis by Bimodal Unit Selection)

THÈSE

pour l'obtention du

Doctorat de l'Université de Lorraine

(spécialité informatique)

présentée par

Utpala MUSTI

Composition du jury

Rapporteurs :	Jean-Claude MARTIN	- Professeur en Informatique, Université Paris-Sud
	Piero COSI	- Senior Researcher, CNR, ISTC, Italie
Examinateurs :	Catherine PELACHAUD	- Directeur de recherche, CNRS-TELECOM ParisTech
	Bernd MÖBIUS	- Professeur, Universität des Saarlandes
	Anne BOYER	- Professeur, Université de Lorraine
	Yves LAPRIE	- Directeur de recherche, CNRS-loria
	Vincent COLOTTE	- Maître de conférences, Université de Lorraine
	Slim OUNI	- Maître de conférences, Université de Lorraine

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria
Laboratoire lorrain de recherche en informatique et ses applications

Mis en page avec la classe thloria.

To my daughter, Samyukta.

Contents

Audio-Visual Speech **7**

Chapter 1 Audio-Visual Speech Synthesis: An Introduction **11**

1.1 **Face modeling and animation** . . . 12
1.2 **Separate visual speech synthesis** . . . 13
1.3 **Simultaneous synthesis of audio-visual speech** . . . 17
1.4 **Conclusion** . . . 18

Chapter 2 Speech Synthesis Using Unit Selection: Literature Survey **19**

2.1 **Unit selection paradigm** . . . 20
2.2 **Segmentation** . . . 22
2.3 **Target cost function** . . . 23
2.3.1 Visual target features . . . 24
2.3.2 Target feature weighting . . . 25
2.3.3 Alternatives to conventional target cost function . . . 27
2.4 **Concatenation cost function** . . . 27
2.5 **Evaluation** . . . 30
2.5.1 Objective automatic evaluation of acoustic and audio-visual speech . . . 31
2.5.2 Human-centered evaluation of acoustic and audio-visual speech . . . 32
2.6 **Conclusion** . . . 35

Chapter 3 Acoustic-Visual Speech Synthesis System: An Overview **37**

3.1 **Corpus preparation** . . . 38
3.1.1 Text selection . . . 38
3.1.2 Acquisition . . . 39
3.1.3 Data processing and parameter extraction . . . 40
3.1.4 Segmentation . . . 43
3.1.5 Bimodal speech database . . . 44
3.2 **Bimodal speech synthesis** . . . 44
3.2.1 Natural language processing . . . 44
3.2.2 Target unit description . . . 46

3.2.3 Bimodal unit selection and concatenation . . . 46
3.3 **Visual speech rendering** . . . 50
3.4 **Conclusion** . . . 50

Chapter 4 Phoneme Classification Based on Facial Data **55**
4.1 **Visual speech segmentation using facial data** . . . 55
4.1.1 Recognition error . . . 57
4.1.2 Forced alignment results . . . 57
4.2 **Learning phoneme kinematics using EMA data** . . . 62
4.2.1 Data acquisition . . . 62
4.2.2 Feature extraction . . . 63
4.2.3 Results . . . 63
4.3 **Conclusion** . . . 65

Chapter 5 Unit Selection **71**
5.1 **Target features** . . . 72
5.2 **Corpus based visual target features** . . . 73
5.2.1 Phonetic category modification . . . 75
5.2.2 Continuous visual target cost function . . . 78
5.2.3 Objective evaluation of synthesis results . . . 80
5.3 **Target feature selection and weight tuning** . . . 83
5.3.1 Unit selection and concatenation . . . 84
5.3.2 Target feature selection and weight tuning . . . 86
5.3.3 Application to AV target cost function tuning . . . 93
5.3.4 Analysis of selected features and their relative importance . . . 94
5.4 **Conclusion** . . . 99

Chapter 6 Evaluation **101**
6.1 **Objective evaluation** . . . 101
6.1.1 Objective evaluation based on comparison of two signals . . . 102
6.1.2 Objective evaluation based on statistical analysis and thresholds . . . 103
6.2 **Human-centered evaluation** . . . 104
6.2.1 Intelligibility tests . . . 104
6.2.2 Quality evaluation tests . . . 105
6.3 **Analysis of perceptual evaluation for better objective metrics** . . . 107
6.4 **Conclusion** . . . 110

Chapter 7 Conclusion **113**

Publications **119**

Appendix A Stimulus for Perceptual and Subjective Evaluation **121**

Appendixs **121**

Bibliography **123**

Audio-Visual Speech

Enacted or animated stories are more popular than audio narrations or those in the books. It is easy to conclude that this is due to its audio-visual nature as it provides a rich experience. Besides entertainment, in general we perceive everything through our ears and eyes, simultaneously. The visual information that is perceived through eyes either compliments or reinforces the auditory information. This applies to speech as well, which is one of the prime modes of communication. Speech perception in the day to day life is primarily bimodal. We see and hear, what is being spoken by people and understand the speech if it is in a known language. Whenever, the auditory input is ambiguous or noise-ridden, we try to supplement the received information by looking at the source, i.e., the speaker. This bimodal nature of speech is illustrated by the observation that, we humans try to have a face-to-face conversation while discussing issues of high importance. This is because, face-to-face communication conveys the complementary information related to speech articulation, emotions, more effectively than just voice. Hence, bimodal speech can be considered more effective in confidence building. Besides entertainment and communication, the basic milestone towards verbal communication, i.e., speech development in babies also has significant contribution of the observation of visual speech along with the corresponding sound (Teinonen et al., 2008; Andersen et al., 1984).

Some of these above mentioned general observations about the advantages of audio-visual speech over acoustic-only speech have been experimentally verified. It has been shown that addition of visual speech enhances speech detection and recognition, thus improving intelligibility when audio is missing, degraded with noise, or where there are multiple sources of speech (Sumby and Pollack, 1954; Ouni et al., 2007; Summerfield, 1979; Schwartz et al., 2004). The evaluation results of visual speech intelligibility by LeGoff et al. (1994) show that the natural face presented 'without' or 'with degraded' audio restores two-thirds of the acoustic intelligibility; with a facial model without a tongue and just a lip model restores half and one-third of it respectively. Speech presented along with facial animation has been observed to be more preferred interface to voice-only presentation. They have been shown to increase the interactive experience of users (Pandzic et al., 1999).

These advantages of audio-visual speech over acoustic speech indicate its vast application possibilities. It has been widely used in entertainment and e-commerce for developing virtual agents. These application do not necessarily need high accuracy of speech articulation. There are other applications which require high accuracy comparable to that of natural audio-visual speech. These include applications for pedagogic activities, for example, virtual language tutors for e-learning, teaching speech articulation to hearing impaired etc (Massaro, 2006). It can also be used to develop virtual announcers for public places that are usually noisy.

Considering all the preceding discussion, it can be said that audio-visual speech synthesis is a significant domain to pursue. But, the advantages of natural bimodal speech can be realized through synthesized audio-visual speech, only if it is comparable to the former. It is so because, humans have implicit expectations from audio-visual speech based on the learning and experience of general face-to-face communications. These are related to temporal alignment and coherence between the acoustic and visual modalities. For instance, while hearing sounds like 'p', we expect a closure of the lips just in time before the onset of that sound. Similarly, we expect to hear high-pitched voice for a conversation where somebody is seen to be in extreme fear. This means that the synthesized audio-visual speech has to have the acoustic and visual streams to be temporally synchronous and coherent with each other.

A majority of approaches for audio-visual (AV) speech synthesis, synthesize the facial animation over speech acoustics, and then perform additional processing for synchronizing the two wherever necessary. This is based on the assumption that AV speech synthesis is a set of two different problems, thereby addressing them sequentially by synthesizing visual speech over synthesized speech acoustics. There are two problems with this approach. To begin with, synchronizing the two streams synthesized separately is not straight-forward. Humans are extremely sensitive to any asynchrony between the audio and speech animation. In fact, this sensitivity to discriminate synchronous speech from asynchronous speech develops very early in humans in their infancy with a significant preference to synchronous speech (Dodd, 1979). Results from (Grant and Greenberg, 2001, 2004) show that human speech perception is extremely sensitive to any lag in the visual domain when compared to audio unlike the other way around. It is also observed that this asynchrony causes a surge in the intelligibility of asynchronous audio-visual speech. Moreover, this also brings in the issue of inconsistency in visual and acoustic domain which might bring in discomfort (Mattheyses et al., 2009). This inconsistency can also affect the final perception of the audio-visual speech, as illustrated by some of the experimental data in (Green and Kuhl, 1989, 1991). These experimental results show that the perception of place and manner of articulation gets affected when inconsistent information is presented in the visual and

acoustic modality. The worst case, where perception of AV speech can be highly affected is that of McGurk effect (McGurk and MacDonald, 1976). In fact, when different facial animation and acoustics are presented synchronously, subjects would experience fusion or combination effect. Fusions effect is seen, for example, when visual/g/ is presented synchronously with acoustic /b/. The result is perceived as /d/. Similarly, when visual /b/ is presented with acoustic /g/) synchronously, it is perceived as /bg/, which is an example of the combination effect. This indicates that synthesizing audio-visual speech by separating the synthesis of the two modalities, might not always ensure the best result in terms of synchrony and coherence of the two modalities. In general, simultaneous processing of acoustic and visual speech is shown to be advantageous with respect to audio-visual integration that are not available with their independent processing (Chen and Rao, 1998).

To ensure a perfect alignment and coherence between acoustic and visual modalities, we advocate synthesizing audio-visual speech simultaneously by treating the two modalities as a single entity. In this thesis, we present our method for audio-visual speech synthesis based on this principle. We base our speech synthesis on the unit selection paradigm. We perform simultaneous synthesis of acoustic and visual modalities by concatenating bimodal units. We keep the natural association between the two modalities intact while doing so, as the visual and acoustic modalities belong to the same speech segment. It should be emphasized that this approach implicitly addresses the above mentioned issues of asynchrony and incoherence. This work can be considered as the crucial first step towards a comprehensive talking-head. Actually, our main focus is to synthesize the audio-visual speech dynamics accurately. The resultant is not a complete talking head yet. Our facial representation is limited to sparse mesh describing the outer surface of the face including the lips. The audio-visual speech does not include the information related to the internal articulators like tongue, teeth and other components necessary for expressive speech. In the course of this work first we studied the bimodal speech corpus, that we acquired, by designing and analyzing visual speech segmentation experiments. Then, we developed the basic system which implemented our idea of bimodal unit concatenation. By using the basic synthesis framework of bimodal unit-selection system, we developed methodologies to improve the bimodal synthesis. In our work, we are addressing the following problems: (1) unit-selection taking both acoustic and visual considerations into account which can drastically increases the complexity, (2) weight tuning, which is a difficult problem in speech synthesis. In fact, we developed corpus specific visual target costs and an iterative target feature weighting algorithm. Finally, we performed perceptual and subjective evaluation experiments through human participants to estimate the intelligibility and quality of our present system.

This thesis is organized as follows. We begin by reviewing the field of audio-visual speech synthesis, in chapter 1. In this chapter, we discuss the ways in which the face has been modeled and animated. We also discuss the various approaches of audio-visual speech synthesis based on separate or joint synthesis of the two modalities. Our speech synthesis system is built on the generic paradigm of unit selection and this is the topic of chapter 2. We review literature related to some aspects of unit selection. It includes, segmentation, that is performed during corpus preparation. Besides, the various building blocks of selection are examined: target description, target and concatenation costs. Finally, we review the ways of evaluating synthesized speech.In chapter 3, we present our work by providing first an overview of our audio-visual speech synthesis system. It also details our audio-visual corpus recording and database preparation for our synthesis system. The resultant audio-visual database that we have is an interesting resource which can be used for studying various phonemes. As a first step in this direction, we have performed segmentation of the visual data. We describe these segmentation experiments, their results and analysis of these results in chapter 4. In chapter 5, we detail different strategies that we developed to optimize our system. It includes designing new visual target features and target feature weighting. Finally in chapter 6, we present the objective evaluation, perceptual evaluation and the analysis done to bring out the relation between the two. We conclude in chapter 7 and explain our future work.

Chapter 1

Audio-Visual Speech Synthesis: An Introduction

In this chapter, we look at some of its earlier synthesis approaches. For any speech, acoustic or audio-visual, to be synthesized from text, the underlying phoneme sequence corresponding to the text has to be first specified. Given this specification, various approaches can be followed for AV speech synthesis. Firstly, these approaches can be divided based on whether the visual and acoustic modalities are synthesized separately or simultaneously. Secondly, the synthesis of acoustic or visual modalities in the case of separate synthesis can be divided based on the synthesis paradigm: rule based, articulatory or concatenative (Theobald, 2007). Thirdly, the approaches can be classified based on their facial rendering technique: 3D modeling of face or image-based.

In a rule-based synthesis system, the well known representative characteristics of speech are simulated using predefined rules. Whereas, articulatory synthesis is done by the simulation of natural process of speech production using models of human anatomy. For instance, air flow is simulated through a controlled model of human vocal tract, and skin of the face is deformed using bones and muscles. Concatenative speech synthesis is performed by concatenating segments of recorded human speech, generally called corpus. This can be put into a broader category called corpus-based speech synthesis which also includes HMM-based speech synthesis. HMM-based synthesis depends on the learning of patterns of speech parameters from a given corpus, which is then used to generate speech parameters. Concatenative approach is like memorizing the whole data, and then accessing the memory at the time of synthesis.

In the following sections, we focus on audio-visual speech synthesis. First, we briefly describe the facial rendering techniques (section 1.1). Then, we discuss the approaches which synthesize the acoustic and visual modalities separately and simultaneously in sections 1.2 and 1.3.

1.1 Face modeling and animation

The face has been encoded and presented in two ways for the purpose of facial animation. The first approach is the 3D modeling of the face. The outer surface of the face is modeled using a mesh of connected polygons. These polygons are made of predefined edges connecting a set of 3D point vertices. Also, changes in the 3D point locations and the consequent changes in the mesh account for the deformations in the face. The first 3D-facial model was developed by Parke (Parke, 1972, 1975, 1982). In this model, the 3D points were defined and controlled by a set of parameters. These parameters were conceptually divided into two distinct sets (functionally they might have an overlap): conformation parameters and expression parameters. The conformation parameters were the ones which define the dimensions of the 3D face. That is, if 3D faces are modeled based on real human subjects for instance, then conformation parameters define the basic 'differentiating' dimensions of that particular human face. These included parameters like aspect ratio of face (height to width), relative sizes specifying forehead, eye separation, nose height, cheek, chin, etc. The expression parameters were those which described mainly the movements of eyes and mouth. They included deformations like jaw rotation, width of the mouth, position of upper lip and corners of the mouth, etc. These deformations might be related to speech or emotional expressions. From these two categories of parameters, the 3D points on the face positions were determined using different types of operations, applied independently to some regions or to the whole face. Eyes were controlled by specific procedures. The other operations included, interpolation, rotation, translation and scaling. The final rendering was done through Phong interpolation (Phong, 1975) based on the parameter specifying the direction of light source. There are many virtual characters which are descendants of this Parke's model (Cohen and Massaro, 1993; Beskow, 1995; Olives et al., 1999). These descendants of Parke's model have various additions to improve the appearance of face and animation: like the addition of the tongue, ears or the back of the head and the addition of control parameters. The advantage of these kind of parametric models is that the whole mesh is specified using a small set of parameters. Parke's parametric model is different from some other parametric models, which are based on modeling the underlying anatomical structure like bones, muscles , skin and forces acting on them (Waters and Terzopoulous, 1990; Waters, 1987; Lee et al., 1995; Ekman and Friesen, 1978). This kind of modeling has been observed to be computationally intensive (Bailly et al., 2003). Some talking heads which present emotional facial animations are based on pseudo-muscle contractions (Cosi et al., 2003; Pelachaud et al., 2001). MPEG-4 standardizes the parametric models by defining a minimum set of 84 feature points (FPs) located on the face. These FPs are controlled by a set of 68 parameters related to perceptible facial deformations

called facial action parameters (FAPs)(Ostermann, 1998).

Besides 3D modeling of the face, the second approach for representing a face is through the usage of facial images. These are most often images of real people. Hence, image-based approaches are generally data-driven. Facial animations using images are generated in two ways. First, it can be done by interpolating few specific images that are representative of the typical articulation of visually identical phonemes called visemes (Ezzat and Poggio, 1998). Alternatively, it can be done by concatenating image sequences (Bregler et al., 1997; E.Cosatto et al., 2000).

The image-based approaches of modeling present more realistic faces. This is because of their proximity to the real facial appearance, which is often described as being photo-realistic. But, this way of encoding or presenting a face is most often limited to a straight-head frontal view of the face. Besides, storage of images incurs significantly higher memory requirement to storage of a few parameter trajectories. On the other hand, 3D-model-based approach is flexible in terms of the view and head orientations in which a face can be rendered. But, an additional processing step is required to add the internal articulators like tongue and teeth to render the complete articulatory information. It is possible to augment the 3D model by adding textural information to make the final facial animation flexible and comparatively photo-realistic Elisie et al. (2001). Another alternative of modeling the face is morphable-models presented in (Cootes et al., 1998; Blanz and Vetter, 1999). These models also embed both geometric and texture related information to present a relatively photo-realistic and flexible facial model.

1.2 Separate visual speech synthesis

Conventionally, AV speech synthesis is considered as two separate problems; the generation of speech acoustics and the generation of facial animation to a given speech acoustics (real or synthesized). Consequently, it has been performed by synthesizing the two modalities separately. Facial animation is generated over a given speech acoustics, which is either synthesized or recorded. This approach requires additional processing to correct the alignment between the two modalities in the case of concatenative visual speech synthesis (Bregler et al., 1997). We refer to the facial animation related to speech as visual speech. We focus on visual speech synthesis stage, considering the acoustic speech already available. Two concepts, which might surface in the discussion of visual speech are: visemes and coarticulation. In the following paragraphs, we first explain these two concepts before going ahead with the synthesis techniques.

Visemes: Visible speech articulation presents similarities for many phonemes. Based on this similarity, phonemes can be divided into different sets. The representative units for each of these

sets are defined as visemes. It is the fundamental unit in the context of visual speech (Fisher, 1968). For example, perception of visual speech while phonemes in the set {p, b, m} are being articulated is almost the same. Hence, they belong to one viseme set. In the current discussion, we mean by viseme, a sequence of visual speech parameters describing a complete segment rather than static targets. On the contrary, we refer to a single sample of these parameters describing a snapshot of a particular target face as 'key frame'. The visual speech parameters can be image frames or trajectories of control parameters or 3D points on the face. This many-to-one mapping of visual speech makes the separation of visual speech synthesis from acoustic speech synthesis advantageous. It is because, the system gets concise due to the reducing in the number of distinct units. In the case of concatenative visual speech synthesis, this increases the possible candidates.

Coarticulation: Coarticulation is the phenomenon in which the articulation of a phoneme is influenced by the articulation of the neighboring phonemes. Synthesized visual speech needs to accurately represent coarticulation. In case of parametric 3D-facial-models, the parameters for animating them have been generated taking coarticulation into account using rules (Beskow, 1995; Pelachaud et al., 1994) or mathematical coarticulation models (Öhman, 1967; Cohen and Massaro, 1993; Cosi et al., 2002). Beskow (1995) mentions that each phoneme has a target vector specifying the typical articulatory gestures. These target vectors are under-specified for some phonemes which are interpolated based on the context to account for coarticulation. Pelachaud et al. (1994) divide phonemes into clusters based on their deformability in different contexts. Phonemes with lower deformability serve as the key frames for coarticulation. Öhman (1967) accounts for the changes during the transformation of a V_1CV_2 (vowel-consonant-vowel) sequence. Cohen and Massaro (1993) implement Löfqvist gestural theory, where phonemes are specified with target feature vectors. Coarticulation is defined as the super-imposition of time-varying dominance functions describing different articulators. These dominance functions are negative exponential functions which peak at the target feature vectors. This coarticulation model has been further augmented by Cosi et al. (2002) by the addition of resistance functions. These resistance functions ensure that some specific target configurations are attained by suppressing the dominance of neighboring phonemes. This is especially important for phonemes like labials and bilabials. Beskow (2004) reports an experimental comparison of various approaches to account for coarticulation. He reports that the mathematical model proposed by Cohen and Massaro (1993) performs well in comparison with the real data; whereas, with respect to intelligibility, rule-based techniques perform better. These models can be optimized through hand-tuning or can be statistically trained using real data acquired using motion capture (Cosi et al., 2002;

Elisie et al., 2001). Ezzat et al. (2002) also perform tuning of a coarticulation model through statistical learning on recorded corpus. Their coarticulation model is similar to that of Cohen and Massaro (1993). Instead of using motion data, they used image-based corpus for tuning their model.

Corpus-based approaches:

Instead of using some explicit coarticulation models, the coarticulation can be implicitly encoded in the synthesized visual speech. This is done in corpus based approaches. Firstly, the complete trajectories of visual speech parameters can be generated using models like HMMs, which are trained on real data (Brand, 1999; Masuko et al., 1998). In this case, the HMM can be modeled as a triphone, which describes a phoneme in the required phonetic context. Alternatively, the complete sequence of visual speech parameters for real motion capture data can be stored and used by concatenating them for synthesis (Minnis and Breen, 2000). In this approach, coarticulation is encoded through the synthesis unit, like triphone or diphone.

In case of concatenative approaches, the visual speech database has to be prepared. Besides acquisition, the corpus needs processing to annotate the individual units in terms of their phonetic labels, segment boundaries, information related to the geometric properties of the faces for ensuring smooth transition at the concatenation points. One of the concatenative approaches for dubbing applications is presented in Bregler et al. (1997). They prepare the visual database by phonetically segmenting an unconstrained video sequence. This segmented video is annotated to include the information based on the orientation of the head, the shape and position of mouth. They use eigenpoints to estimate the fiduciary points on the face (mouth, teeth, chin and jaw line) using 26 hand annoted images. Also, the synthesis is done by the concatenation of triphone video clips. The synthesized mouth sequences are then morphed onto the background video sequence. The resulting video sequence is compressed or stretched to time-align with the target audio between phoneme boundaries.

The synthesis described in (E.Cosatto et al., 2000) is based on the concatenation of variable length video sequences of mouth images (and also other facial parts). The database is described in terms of 3D geometric features of the head and appearance features extracted by Principal Component Analysis (PCA). They further subdivide the facial parts into cheeks, teeth, tongue, jaw, etc to make the synthesis more flexible. The final synthesis is done by overlaying bitmaps of the facial parts present in the database onto a background video as in (Cosatto and Graf, 1998). There are other similar works of image based concatenative approaches (Weissenfeld et al., 2005; Liu and Ostermann, 2009). For instance, Weissenfeld et al. (2005) use Locally Linear

Embedding (LLE) to describe the appearance parameters of the mouth images unlike Cosatto and Graf (1998) who use PCA. Liu and Ostermann (2009) use PCA to extract appearance parameters and Active Appearance Models (AAM) to extract the geometric parameters of the face (lip width, lip height, visibility of teeth and tongue). A similar approach, but which is based on parametric 3D facial model is presented in (Ma et al., 2006). In this approach, the control parameters extracted from recorded 3D facial marker data are concatenated using unit selection. The resultant trajectories are used to animate virtual conversational agents.

Some concatenative approaches combine HMM and concatenative approaches for visual speech synthesis. One such work is presented in (Lijuan et al., 2010). It is image-based approach where the selection process is guided by the trajectory of lip movements generated by trained HMMs. These HMMs are trained by the AV-speech corpus. This approach is similar to an earlier work by Govokhina et al. (2006). In that, phonetically aligned trajectories of 3D facial markers are selected based on the trajectories generated by trained HMMs. A hybrid unit selection and HMM based approach for visual speech synthesis is presented in (Edge et al., 2009). This work uses the selected units to train state-based models and search through these learned models through Viterbi type algorithm. The similarity in speech acoustics (acoustic parameters) was used to guide through unit selection. The final sequence of state-based models is used to generate smooth trajectories for visual speech. Bailly et al. (2009) describe a system which generates articulatory gestures (control parameters) for a video realistic (image based) facial animation using HMMs. They incorporate a phasing model to learn the lag between visual gestures and corresponding speech acoustics. They compare this HMM-based technique which includes the phasing model with 3 other techniques: (1) concatenation of articulatory gestures selected based on the phonetic context, (2) concatenation of articulatory gestures based on selection that is guided through the phasing model based HMM, (3) trajectory generated by HMM models trained on audio-synchronized articulatory gestures. They conclude that the phasing model based HMMs improve the synthesis.

Almost all of these works report lip-synchronization problems. Bregler et al. (1997) report that plosives were observed to have occasional lip-synchronization problem, Cosatto and Graf (2000), report lip-synchronization being criticized in subjective evaluation. Geiger et al. (2003) present the perceptual evaluation of the synthesis approach presented in (Ezzat et al., 2002). They report that the synthesized audio-visual speech is not comparable to the natural audio-visual speech, to the extent that is required for developing applications for teaching language or speech articulation to the hearing-impaired.

1.3 Simultaneous synthesis of audio-visual speech

The potential application of audio-visual speech hinges not only on the accuracy of the synthesized visual speech, but also on the extent to which the acoustic and visual streams agree with each other in terms of synchrony and coherence. It is obvious from the previous section that, through the separate synthesis of acoustic and visual modalities, these conditions are not always guaranteed. In this section, we look at approaches which synthesize audio and visual speech simultaneously. The central mechanism of all these approaches is to keep the association between the visual and acoustic modalities, thereby preserving the natural synchrony and coherence. Majority of approaches in this category are based on the concatenation of synchronous bimodal units. One approach presented by Tamura et al. (1999), uses HMM models trained using synchronous audio-visual speech data to generate bimodal speech parameters. But, it should be said that this approach was still at a much preliminary level as the generated visual speech parameters were related only to the lip contours.

The concept of synchronous bimodal unit concatenation for Swedish AV speech synthesis has been presented in (Hallgren and Lyberg, 1998). The visual speech information is recorded as trajectories of 3D markers all over the face, especially around the lips. The recorded marker information is used to control a 3D model of the head. This head model is further textured to make it look more natural.

Two recent image-based approaches that use concatenation of bimodal units are (Fagel, 2006; Mattheyses et al., 2009). In (Fagel, 2006), AV speech synthesis is done for German by concatenating synchronous bimodal polyphone segments. This was with a 4-minute corpus consisting of bimodal speech: video of speech aligned with the corresponding phonetic transcript. The selection of polyphone segments for concatenation was based on a concatenation cost calculated as a weighted sum of acoustic and visual concatenation costs. The pre-selection of possible polyphone segments from the corpus is based on chunks (longest polyphone segments that are available in the corpus), and the visual joint cost calculation is based on the pixel to pixel color differences in the end frames of the video clips to be concatenated. Hence, it is quite clear that synthesis incurs a large overall processing time. In (Mattheyses et al., 2009), the conventional unit-selection technique which has been widely used for acoustic speech synthesis is extended to perform AV speech synthesis. It is done by including an additional join cost term for visual join discontinuities. Their system is similar to the one explained in (Liu and Ostermann, 2009) in terms of the visual features extracted and used to describe the facial geometry and appearance. These methods like any image-based technique incur high storage requirement when compared to a 3D-model based approach.

1.4 Conclusion

In this chapter, we have discussed various techniques to model the face that are based on either its 3D or image-based representation. We have also discussed the various pros and cons of each technique. Further, we have also examined some approaches of AV speech synthesis that are based on either the sequential (synthesizing facial animation after acoustic speech synthesis) or simultaneous synthesis of the two modalities. We have highlighted the disadvantages of the former. Consequently, we are in favor of synchronous, data-driven synthesis of audio-visual speech. Our approach is based on this line of synthesis. As can be seen in chapter 3, our approach is using a unit-selection paradigm to synthesize both visual and acoustic modalities simultaneously. In the following chapter, we present a survey of various aspects of unit selection and then we introduce our system in chapter 3.

Chapter 2

Speech Synthesis Using Unit Selection: Literature Survey

Speech synthesis is a well established field of research with significant progress in the past three decades. Though synthesized speech is getting closer to human speech, it is still far from being considered a solved problem. In addition, we are still away from a perfect all-purpose speech synthesizer. This is true for both acoustic-only and audio-visual speech. Among the synthesis techniques concatenative techniques have become very popular in recent times. These methods have been widely used and evolved for acoustic synthesis. Nevertheless, the paradigm is generic and has been extended to visual or audio-visual speech synthesis. In the earlier concatenative acoustic synthesis, fewer instances of each diphone were stored in the inventory. The synthesis specification included the prosodic description related to duration and pitch of targets in the sentence to be synthesized. At the time of synthesis, these diphones were modified using signal processing techniques to bring in the changes related to prosody and then concatenated. This kind of intensive signal processing done on the waveform distorts its naturalness. The advantage of this system was the small size of the diphone inventory which was a necessary requirement at the time of its usage. Moreover, it can be said that in spite of usage of signal processing, it does not account for all the variations of speech accurately.

As computer storage is getting cheaper and faster, it has become possible to store huge speech database many times larger than the earlier smaller inventory of diphones. Usage of a huge corpus, makes it possible to include a large set of candidate diphones with large variability in their waveforms. Moreover, it is even possible to have longer synthesis units than a diphone. In fact, it is even possible to search for whole sentences or big chunks of sentences. This indicates the drastic reduction in the need to process the speech signal. Consequently, the resultant speech preserves the naturalness of the original speech as the speech segments are concatenated with

little to no signal processing.

Nevertheless, the usage of a large speech corpus has different problems. A large variance in the synthesis candidates means that selection has to be done carefully, to synthesize speech which is similar to a natural utterance. This is the classical unit selection problem. We discuss some of the issues of unit selection techniques, and the approaches that have been applied to resolve them. In the following sections, we first give a brief introduction of the emergence of the framework of unit selection and its basic paradigm (in section 2.1). In section 2.2 we give a short description of the segmentation techniques used in corpus preparation, then a description of pre-selection of candidates and the conventional target cost formulation based on independent feature space assumption and its tuning (in section 2.3). Next, (in section 2.4) we give a brief account of the ways join evaluation techniques have been analyzed for their correlation with human perception of discontinuity when non-contiguous units are concatenated. Finally, (in section 2.5), we deal with the objective and perceptual evaluation methodologies that are generally employed to estimate and sometimes qualify a text-to-speech synthesis (acoustic or audio-visual) for its use in a specific domain.

2.1 Unit selection paradigm

Unit selection depends on the selection of the best possible set of units from different variants available in the corpus. Consequently, the first requirement is to have a corpus that not only has a good coverage of the possible speech variants, but which is also comparatively small to keep the search time short (Möbius, 2000). Given a particular speech corpus, the quality of the synthesized speech using unit selection depends on its usage. Many factors affect the synthesis results. For example, concatenation of units can be said to be the most obvious reason for audible disruption and many initial systems were based on the reduction of concatenation points (Sagisaka, 1988). In (Sagisaka, 1988), the selection of longest segments is given preference and the concatenation at certain locations like at CV (consonant-vowel) boundaries or in the middle of vowels is penalized. Alternatively, when it is not possible to avoid concatenation of non-contiguous units, minimization of distortion at the concatenation point minimizes the quality degradation (Takeda et al., 1990; Iwahashi et al., 1992). Besides reducing the concatenation of non-contiguous units, there are other necessary factors that need to be considered. For example, the phonetic context of the selected unit and the speech realization of the unit itself seems important (Takeda et al., 1990; Iwahashi et al., 1992).

The search procedure proposed in (Hunt and Black, 1996) for unit selection offers a unification framework where all the above mentioned considerations can be included while determining a

possible optimal solution to the selection-concatenation problem. For a sequence of candidates u, and a sequence of required target units t; the paradigm presented by Hunt and Black (1996) optimizes a total cost function which is a weighted sum of the following:

- The perceptual suitability of u, for t, which is called the target cost, denoted by $TC(t, c)$.
- The total discontinuity at all the concatenation points, called the join cost denoted by $JC(c)$.

Denoting the weights of the target cost and the join cost by w_{tc} and w_{jc} respectively; from a given corpus, the search for the final sequence of candidates is done based on the optimum candidate sequence which minimizes the total cost (C) as shown below:

$$C = \min_{u} w_{tc} TC(t, u) + w_{jc} JC(u) \tag{2.1}$$

Here, the pre-selection of units is based on a same-size units like phones or diphones for each target position. This pre-selection is based on the target cost determining the suitability of the candidate and its context. Also, in this general framework, the selection of longest contiguous candidates is enforced implicitly by making the individual join costs for any two contiguous units in the corpus zero (Balestri et al., 1999). This has the advantage of taking into account the variability of speech realization besides reducing the concatenation artifacts for the selection of possible best set of candidates. In contrast, some methods explicitly search for longest contiguous units for concatenation called non-uniform unit selection, where the units sought for concatenation are not of same size or type (Taylor and Black, 1999; Boëffard, 2001; Schweitzer et al., 2003). This is different from the earlier paradigm which is implicitly non-uniform unit selection, as there might be many contiguous segments of variable size in the final synthesized speech. Clark et al. (2004) give a good description of the practical aspects of building a unit selection based speech synthesizer. Taylor (2009), gives a comprehensive overview of the different approaches addressing various aspects of unit selection based speech synthesis. Our approach is based on the first paradigm, which is an implicit non-uniform unit selection.

Extending unit selection to audio-visual speech synthesis

In majority of AV speech synthesis approaches, visual speech is synthesized over an available acoustic speech that is either synthesized or real. In the case of visual or audio-visual speech synthesis using unit selection, the selection of segments has to be done considering the requirements of visual modality also. This involves the inclusion of visual criteria during pre-selection, i.e.,

in the target cost function, and also additional join criteria to account for the visual modality related discontinuities in the join cost function.

2.2 Segmentation

It is obvious that unit selection depends on a speech database. Segmentation is one of the steps of this database preparation, in which recorded speech is divided into phonetic segments by demarcating their temporal boundaries. These phonetic segments constitute the basic building blocks for synthesis. Speech segmentation without any other specifier is conventionally used to refer to acoustic speech segmentation. Though the best way in terms of accuracy is manual segmentation (Cosi et al., 1991; Ljolje and Riley, 1993; Ljolje et al., 1997), it is time-consuming, laborious and hence costly. For this reason, automatic speech segmentation is considered a good alternative. The most popular and widely used technique for automatic speech segmentation is to force a HMM based phonetic speech recognizer to recognize the speech to a given phonetic transcript. Demarcation of phonetic boundaries is a result of this forced-recognition which is conventionally called forced alignment. This alignment technique has avoided the need for manual alignment to some extent and also considered good enough for HMM training that is required in speech recognition. But, segmentation needs to be more accurate for concatenative speech synthesis especially for those which are based on concatenation at phoneme boundaries. Consequently, various methods have been used for the refinement of the phonetic segment boundaries further (Toledano et al., 2003). Some of the recent works use a combination of segmentation methods to derive multiple time marks to arrive at more accurate segmentation (Kominek and Black, 2004; Park and Kim, 2007).

For concatenative visual or AV speech synthesis, generally the boundary time-marks determined by the acoustic speech segmentation of an audio-visual corpus are used while defining the candidates in the corpus (Bregler et al., 1997; Hallgren and Lyberg, 1998; E.Cosatto et al., 2000). This way of segmentation is widely followed and practically shown to work for visual speech synthesis. Nevertheless, this is not in accordance with the underlying principal of speech production. The speech articulators have to be ready with a target configurations required for the production of a sound (phone) for it to happen. That is, the start and end in the visual and acoustic modalities may not necessarily be the same. Some works have tried to learn this time lag between acoustic and visual by adding phasing models (Govokhina et al., 2007; Bailly et al., 2009). These phasing models are arrived at through iterative process involving HMM learning, forced alignment of trajectories of articulatory gestures, comparison with the acoustic segment boundaries and adjustment of visual segment boundaries. Since, speech segmentation

works through recognition of the speech segment, it provides an interesting tool to study the unique characteristics of phonemes. We exploit this idea to characterize phonemes (Chapter 4).

2.3 Target cost function

Measuring the suitability of a candidate in the corpus for a target position in the speech to be synthesized is a necessary step in unit selection. The efficiency of a target cost function in ranking and pre-selecting candidates also affects the probability of a good join and thus the quality of the synthesized speech. Generally, the target and the candidate are defined in terms of factors which are known to account for the variation in speech realization based on phonetic and linguistic studies. These factors are at the abstract level which are not directly expressible in terms of the actual speech parameters quantitatively. These are referred to as high-level features. These features can take either non-negative integral values or can be categorical. These features might include:

- Phonetic features like the phonemic identity of the current unit and the neighboring units (context), type of phoneme (vowel, consonant), voicing of phoneme (voiced, unvoiced), manner of articulation etc.

- Linguistic features like position of a syllable at various levels (word, rhythm group, sentence, etc); position of word in a rhythm group or sentence; type of sentence etc. These features generally account for the various suprasegmental prosodic patterns. Some of the features in this category might be language specific.

Target feature set can also include features that are based on the statistical analysis of speech related parameters which are extracted from corpus, which are referred to as low-level features. For example, some systems use prosody prediction models that mainly provide duration and pitch specification of the segments to be selected. These prosody prediction models are trained on real corpus. It helps in reducing the number of high-level target features needed to describe prosody (Latacz et al., 2010). The low-level target features are also used to speed-up the pre-selection by reducing the search space (Black and Taylor, 1997).

Lot of systems use target feature set which consists of majority of higher level features (Hunt and Black, 1996; Coorman et al., 2000; Latacz et al., 2010). Some systems use higher-level target features exclusively to allow the automatic selection of candidates with suitable prosodic characteristics rather than prediction based on prosodic models (Prudon and d'Alessandro, 2001; Colotte and Beaufort, 2005). The target cost is generally calculated as a weighted sum of the

individual feature costs. Three kinds of target feature costs have been generally used (Coorman et al., 2000):

1. Categorical distance measures: Where the distance is either a binary valued or non-negative integer-valued function between categorical features.
2. Scalar distance measures: Non-negative real valued function for features like duration, F0 etc.
3. Vector distance measures: Distance calculation for multi-dimensional features, like the acoustic and visual feature vectors.

Categorical distance measures are calculated for the high-level target features while the other two are based on the low-level features. For AV speech synthesis the set of target features has to be augmented to include the information regarding speech realization in the visual modality. Besides the target feature description, the weighting of features for a given target set in the order of their relative importance is crucial for selection. These aspects are presented in the following two sections. Besides the conventional target cost, alternatives have been proposed which we review in subsection 2.3.3.

2.3.1 Visual target features

For the visual speech synthesis many of the high level target features used are those which describe the visual or audio-visual target. These features might include typical articulatory characteristics like lip closures in bilabials. They might also include rate of speech related characteristics. Besides features which are equally important for visual and acoustic speech realization (e.g., place of articulation), or those which account more for the acoustic realization (e.g., voicing), there are some features which are more important for describing a visual target (e.g., shape of the lips during the articulation of a phoneme). Many of the concatenative AV speech synthesis systems use a visual target cost based on the similarity of two phonemes in terms of visible facial deformations, as described below.

In (Bregler et al., 1997), a categorical phoneme context distance is used for the selection of triphone which accounts for the visual target cost. Phonemes of same label are assigned 0 cost, and phonemes belonging to two different viseme classes are assigned 1, and different phonemes of same viseme class are assigned a cost between 0 and 1 which are derived from confusion matrices described in (Owens and Blazek, 1985).

In (E.Cosatto et al., 2000), a viseme distance matrix is used for the calculation of target cost between a target and candidate frame. It is calculated based on the similarities in the visual

domain irrespective of the differences in the acoustic domain. The selection of the visual segment is based on duration and phonetic label of the target segment which is obtained from the acoustic speech. Each target frame is specified in terms of the phonetic annotation of a window of frame sequences consisting of some fixed number including itself to account for context. The window length is different for each phoneme. The candidate is selected with the most proximate context which is measured by the target cost. The target cost weight vector is based on the exponential decaying influence inspired by (Cohen and Massaro, 1993). Weissenfeld et al. (2005) use a similar visual target cost where the difference matrix is calculated based on the visual difference matrix populated using the Euclidean distance in visual feature space. It is based on the assumption that each phoneme can be described by its mean visual feature vector, which is speaker and corpus specific. In (Mattheyses et al., 2010), a similar visual target cost calculated based on corpus is included. The difference matrix that is calculated represents the inter-phoneme visual distances based on the mean and variance of visual parameters at the middle of the phoneme units present in the corpus. These kind of cost functions which are calculated for a specific corpus don't guarantee optimum performance for any other corpus in general.

2.3.2 Target feature weighting

The target cost tuning involves the determination of relative importance of target features and assigning weights to the individual target feature costs to be used for target cost calculation. Ideally, it is done in such a way that the ordering of candidates based on the target cost corresponds to their perceptual suitability as a target. Since the synthesized speech has to be at least acceptable, intelligible and near natural speech for human listeners, some system tuning techniques are based on human listening tests (Coorman et al., 2000; Alías et al., 2004). Listening tests are time-taking and require human subjects which make them practically costly. Moreover the scope of this kind of tuning is limited to a few set of sentences and hence it cannot guarantee consistent synthesis results. It becomes further difficult when the set of target features is large. Hence automatic weight tuning has been applied in many of the works (Hunt and Black, 1996; Meron and Hiros, 1999; Park et al., 2003; Alías and Llorà, 2003; Colotte and Beaufort, 2005; Latacz et al., 2010).

The target feature weighting techniques can be divided into two categories: (1) joint weight tuning of concatenation and target feature cost functions, either at the individual unit level selection by using pairs of synthesis units or at sentence level, (2) separate weight tuning of target and concatenation cost functions, generally by tuning the target feature costs at the synthesis unit or phonetic segment level. In both the techniques, a real segment or sentence not

included for selection is treated as the target, and selected or synthesized from the corpus. The target and the selected units are compared using objective distance measures to perform the tuning.

One of the two techniques presented by Hunt and Black (1996) called 'weight space search' (WSS) is based on the first category of weight tuning. It is based on the usage of targets from real sentences held out for training from the synthesis database. The weight tuning is done by searching the weight space, in such a way that the waveforms of synthesized sentences and that of real sentences are similar. The weight space search is limited to a finite set of weight combinations and choose the best weights among the searched combinations for defining the target cost function. This method is computationally very expensive in case of large number of features and possible set of target feature cost values. Meron and Hiros (1999) presented acceleration techniques for WSS by partial synthesis and comparison. Alías and Llorà (2003) performed target tuning by using genetic algorithm for doing the weight space search. The advantage of this is that the search space is randomized and search evolves towards better weight combination, unlike in the former works where a fixed finite combinations were searched. Latacz et al. (2010) also present an automatic weighting technique for tuning target features and concatenation costs together. In their technique the ordering given by weighted sum of target cost and concatenation cost, and the ordering given by an acoustic distance metric are compared. A selected error is calculated based on the mismatch in this ordering. They refer this technique as Minimum Selection Error training. Further, they propose that the set of weights obtained for all the candidates treated as targets being clustered using decision trees.

One of the techniques which performs target feature weighting separate from concatenation costs weighting is based on multiple linear regression (Hunt and Black, 1996). Using this method, the target feature weights for each phoneme in a language's phoneme set are tuned separately to come up with different target costs for different phonemes. Each of the candidate in the database is considered as a target each time and the n most similar candidates are selected from the phoneme's candidate set leaving the target out. The ordering of candidates for the pre-selection of n candidates is based on an objective distance measure. The target weights are determined using Linear Regression such that the target cost predicts the objective distance measure. Meron and Hiros (1999) presented a way to extend this regression training (RT) for weighting the target features and concatenation costs together using target pairs unlike single targets. They also propose clustering of phonetic contexts by using a decision tree to split the phoneme pairs into different clusters. This is done with a phonetic contextual question which split the phoneme pairs into sets with least regression error at each level (using RL).

Each target feature accounts for variations in speech, and their duration. Based on the discriminative information accounted by each of the features, they have been weighted in Colotte and Beaufort (2005). Acoustic representation of a particular phoneme units were divided into clusters through K-Means algorithm using Kullback-Leibler divergence as the similarity index. The weight of the feature is based on its discriminative information between the different clusters. This is applied to all the phonemes in the phoneme set of the language separately. Another approach to weight tuning is to view unit selection as a classification problem (Park et al., 2003), in which instead of defining an objective function to account for the subjective speech quality, the classification error is taken as the objective function to be optimized. It is difficult to compare these methods in terms of their synthesis results. There are many factors which vary in these approaches, like, speech corpus, test sentences, evaluation methodologies etc. Hence, it is not straight forward to relatively judge their performance.

2.3.3 Alternatives to conventional target cost function

The target cost put forth by (Hunt and Black, 1996) was weighted sum of individual feature costs (differences). Whenever a candidate with the exact target feature description is not available, the candidate selected for synthesis based on this simple formulation for measuring target-candidate similarity or rather dissimilarity might not always reflect the actual human perception. The following two cases need little more consideration: (1) where a candidate with required exact feature description is not available, but, a candidate with a speech realization similar to the required one but with a different feature description is available; (2) where neither the a candidate with exact feature description nor with a similar speech realization is available, in which case, a better possible alternative(s) have to be selected. To consider the speech realization besides the target combination alone of candidates, alternate approaches for target cost calculation have been proposed which base the selection on the perceptual similarity estimated through acoustic distances (Taylor, 2006). The main idea behind the proposed method is to have representation of the segment to be selected in terms of the low-level features by using the high-level features. This was done by clustering the candidates of a particular phoneme using acoustic distances and using decision trees to choose a cluster for unit selections by Taylor (2006).

2.4 Concatenation cost function

It is known that the acoustic speech quality degrades due to the concatenation of non-contiguous speech segments. Also, studies have shown that considering the spectral smoothness at the concatenation point improve the naturalness and intelligibility (Takeda et al., 1990; Iwahashi

et al., 1992). This holds for visual speech as well. Hence, any abrupt jump in the visual speech sequence can create perceptual discomfort and confusion. Consequently, the focus on reduction of concatenation artifacts arguably dates back to the onset of concatenative speech synthesis itself. Especially in unit selection based speech synthesis, there is a wide variability in the candidates for each target required. This results in a large variance in the concatenation points as well, like in the middle of a phone when diphone is the synthesis unit. Good concatenation is important not only for a good synthesis quality, but also for intelligibility (Clark et al., 2007).

While designing good concatenation strategies for unit selection, different approaches have been followed. The candidate preference for concatenation is based on the observation that naturally contiguous units automatically join well. Hence, all systems give preference to contiguous units in the corpus, besides considering important phonetic and prosodic characteristics. In fact, some systems go further and search the longest possible units from the corpus, so as to reduce the number of concatenation points (Schweitzer et al., 2003). Since it is infeasible to have a naturally contiguous speech in the corpus for every target sequence to be synthesized, various join optimization techniques have been developed.

The most widely followed approach for concatenation is to minimize the differences at the concatenation points. This strategy is based on the observation that huge differences in the waveforms at the concatenation points account for perceptible degradation. Various distance metrics calculated using various acoustic parameters have been explored for estimating the perceptual degradation due to joins. Cepstra, line spectral frequencies, log area ratios, mel frequency cepstral coefficients, multiple centroid analysis (MCA) coefficients, linear predictive coding coefficients are a few of them. Euclidean, Absolute, Kullback-Leibler, Mahalanobis are some of the distance measures explored. Given these many alternatives, it becomes necessary to base the join difference estimation using those measures that correlated well with human perception. Hence, there are many attempts to evaluate the parameter and distance measure combinations to rank them based on their correlation to human perception of join discontinuity. Some of these works ask listeners to evaluate joins on a 5-point MOS scale and compare these scores with the distances calculated using various metrics and acoustic parameters (Wouters and Macon, 1998, Vepa et al., 2002, 2004, Donovan, 2001, Bellegarda, 2004). In some other works, the comparison between human perception and distance metrics is based on the detection of a join, i.e. a binary score (Klabbers and Veldhuis, 1998,2001, Stylianou and Syrdal, 2001, Pantazis et al., 2005). The results presented in the various works don't agree much with each other. Kullback-Leibler divergence has been reported to perform well with different parameters in some of the works (Klabbers and Veldhuis, 1998; Donovan, 2001; Vepa et al., 2002). The highest correlation

reported between the objective distance measures and the perceptual evaluation results is 0.66 which has been deemed low. Hence, the choice of any particular speech parameterization and a distance measure does not ensure an accurate estimate of perceptual disruption at the join.

While trying to reduce the join disruption due to concatenation, naturally contiguous units can be used to determine the set of units which can naturally join well. This can be based on their proximity to naturally good joins, i.e., contiguous units in the corpus. The work done by Vepa and King (2003) can be considered to be in this direction. In their work, the natural evolution patterns in the acoustic parameters are learned from the corpus, and used as the basis for the evaluation of a join and defining a join cost function. Naturally contiguous speech samples are never perceived as discontinuous, though they are seldom exactly the same. From this observation, it can be concluded that humans are insensitive to a slight disruption at the concatenation point. This has been used as a basis for formulation of the evaluation of joins by Coorman et al. (2000). They have described a masking function to evaluate a join . Consequently, below a certain transparency threshold the join cost is zero.

Irrespective of the distance between two concatenation points, it has been observed that join disruption is not perceived uniformly across all the phonetic contexts. In other words, the perceptual degradation of speech is high in some phonetic units and contexts than some others. Syrdal, 2001, 2005 report a systematic study of the human sensitivity to disruption at various contexts, a summary of the results presented is as follows: discontinuities are perceived more with female voice based speech synthesis to male voice based speech synthesis, higher in vowels than in consonants, higher in diphthongs than to other vowels and higher in sonorant phonemes than non-sonorants. They also reported a comprehensive list of join discontinuity detection (%) based on the phoneme type. This shows that phonemic context is important and concatenation in certain contexts or phonemes are less preferable to some others and hence phoneme independent handling of concatenation strategies might not be the best.

Concatenation of audio-visual units

All the salient points considered for acoustic unit concatenation are equally applicable for visual or audio-visual unit concatenation. Here, the way the distances are calculated for units at concatenation points depends on the visual features. For example, in (Bregler et al., 1997), a distance to measure the difference in lip shapes in the overlapping segments of adjacent triphones is included to account for the concatenation cost. It is calculated as the Euclidean distance (frame-by-frame) between four element feature vector of articulatory features, outer-lip-width, outer-lip-height, inner-lip-height and height of visible teeth. The place of concatenation is de-

cided based on the place of least difference in the lip shapes. In (E.Cosatto et al., 2000), the visual concatenation cost has two components, the skip cost and a transition cost. Skip cost is a penalty for any two frames which are not contiguous in the corpus and calculated based on the ordering of frames in the corpus, 0 for any two naturally contiguous units or frames . The transition cost is calculated based on the visual distance between two frames. Its calculated as the Euclidean distance of two PCA feature vectors extracted based on the appearance. Similarly, in (Ma et al., 2006), two frames are given zero concatenation cost when they are contiguous in the original corpus, for those frames which are not contiguous its calculated as a sum of a minimum constant value and a variable component calculated based on the frames. The variable component in turn has two components, one of which is calculated based on the distance calculated between the two frames. The second component of this variable concatenation cost ensures that the visemic transition in the synthesized and original corpus are the same. For example two frames i and j can be concatenated if the preceding frame of j belongs to the same visemic label as that of i. The trajectories at the joins are made smooth by applying a low pass filter and cubic splines. In (Fagel, 2006), the video joint cost calculation is based on the pixel to pixel color differences in the border frames in the segments to be concatenated (computationally expensive).

2.5 Evaluation

We have considered various aspects of unit-selection based speech synthesis. In this section, we present the ways of evaluating synthesized speech. This is necessary for exploring different approaches to improve synthesis quality, in which case changes need to be quantified and for comparative evaluation of different synthesis systems. These can be related to selection, concatenation and overall system tuning. As synthesized speech is targeted for human perception, the most accurate way to evaluate a synthesized speech is perceptual evaluation by human subjects. In-spite of its accuracy, automatic evaluation is often done instead, by comparing synthesized speech with a reference speech. This reference is generally recorded real speech which is not included in the corpus. This comparison is quantified using some objective evaluation metrics. In the following, we present the objective evaluation metrics and then the perceptual evaluation by human subjects. The evaluation of synthesized speech by human subjects is done in two fronts: subjective evaluation of quality, and perceptual evaluation of intelligibility.

2.5.1 Objective automatic evaluation of acoustic and audio-visual speech

Various distance measures have been proposed for comparing real and synthesized speech signals. For example, cepstral distance is used as a distance measure in many works for acoustic speech (Hunt and Black, 1996; Meron and Hiros, 1999; Alías and Llorà, 2003). (Latacz et al., 2010) used constituent distances measures for duration, f0 and spectrum. Objective evaluation of audio-visual speech is generally done based on an independent objective evaluation of visual and acoustic modalities. Alternatively, the objective evaluation of only one modality is performed sometimes, based on the focus of analysis. For instance, in (Huang et al., 2002) only the synthesized visual speech is evaluated. It was done using three objective evaluation metrics . These were developed for estimating the precision (naturalness) and smoothness of visual speech; and synchronization between acoustic and visual modality. Firstly, precision was estimated using the sum of Euclidean distance between the real and synthesized sentences, calculated on visual parameters. Secondly, smoothness was estimated using the sum of Euclidean distance calculated between adjacent frames in the synthesized speech which are from non-contiguous locations in the corpus. Lastly, audio-visual synchronization was estimated based on the phonetic labels of synthesized frames. For this, only a few important phonemes were considered, which belong to one of the following two categories. The first category was of those phonemes which have a change in the direction of the mouth movement, i.e., from closing to opening or vice versa. The second category included those phonemes which have maximal mouth shapes like open or closed mouths. Similarly Euclidean distance measure has been used by some others (Weissenfeld et al., 2005).

Instead of comparing real and synthesized speech, Liu and Ostermann (2009) use average target cost, average segment length and average visual difference between frames as the objective evaluation metrics and minimize them during total cost tuning. This is based on the assumption that the average target cost is representative of the lip-synchronization (audio-visual synchronization) and the other two metrics represent the smoothness of the speech animation. But finally, for evaluating the weights resulting from the tuning process, cross correlation coefficient between the PCA coefficients of the synthesized and real sentences was calculated to represent the subjective quality of the synthesized visual speech. Similarly, (Bailly et al., 2009) report the comparison of different articulatory gesture prediction techniques using the correlation coefficient between original and predicted gestures. For objective evaluation of the synthesized visual speech, Ma et al. (2006) use average errors of normalized articulatory parameters (lip-height, lip-width, lip-protrusion) between the original and synthesized speech. Though these techniques present a fast way to estimate the dissimilarity between two speech realizations, their correlation

with human perception has not been quantified systematically.

2.5.2 Human-centered evaluation of acoustic and audio-visual speech

For any text-to-speech synthesis system, an evaluation of the overall system performance by human subjects is inevitable irrespective of which domain it is to be deployed. This is so because, the final users of any synthesized speech are humans. Manual evaluation of text to speech synthesis system is generally done to evaluate at least two aspects of its synthesized speech: quality (especially naturalness) and intelligibility. These specific aspects to be evaluated and the evaluation techniques depend on the target applications. These possible applications can be, conversational agents for hearing impaired, or for movie dubbing with different audio or video track, human-computer interaction to mention just a few to name. Some of the aspects which are application specific are the following: (1) suitability of the speaker which depends on his/her voice clarity, ethnicity and native language which affect pronunciation and also pleasantness for e-commerce related application, (2) time required for synthesis, (3) prosodic component accuracy, (4) overall intelligibility.

Generally, the quality of synthesized speech is evaluated in terms of the subjective evaluation measures, Mean Opinion Score (MOS) or DMOS (degradation Mean Opinion score). These are also know as Absolute Category Rating (ACR) or Degradation Category Rating (DCR) respectively. In these evaluations, human subjects are generally asked to give a categorical score with respect to some particular aspect of the speech whether it be acoustic, visual or audio-visual speech. The difference between the two (MOS and DMOS) is that in the second case, the score is generally given with respect to a reference, generally the real utterance. The different aspects of quality can be broadly classified into naturalness, pronunciation, pleasantness, overall comprehension and intelligibility. Their different categories depend on the attribute that is being evaluated. The different aspects to be evaluated also depend on the method used for face modeling and rendering, besides the target application domain. For example, for a human-computer interacting experience like virtual avatar for e-commerce, the likability of the virtual character and its expressiveness of emotions are also important for confidence building. For example, in (Ma et al., 2006) the accuracy and naturalness of the synthesized speech are reported in comparison with that of natural audio-visual speech using the usual 5 point MOS scale. Similarly, Bailly et al. (2009) report subjective evaluation of audio-visual speech by synthesized image sequence over natural audio by preference tests based on 5-scale MOS test (5-very good, 4-good, 3-average, 2-insufficient, 1-very insufficient). Alternatively, naturalness tests are conducted asking the listeners to identify sentences as real or synthesized instead of MOS rating, which are

called Turing tests (Geiger et al., 2003; Liu and Ostermann, 2009).

The evaluation of intelligibility is done by the perceptual evaluation at various levels, phoneme, word and sentence. For phoneme level intelligibility testing, rhyme tests and nonsense words are utilized. In rhyme tests, words differing in a single phoneme segment are presented and asked to report the actual word that is heard by a human subject. Diagnostic Rhyme Test (Fairbanks, 1958), Modified Rhyme Test (MRT) (House et al., 1963) are two of the well known rhyme tests. Both use single syllabic word sets, former consists of word pairs, whereas the later has sets of six words each. Sentence-level tests are conducted to assay the intelligibility of words in context. The most commonly used test is with semantically unpredictable sentences (SUS) proposed by Benoît et al. (1996). In these tests special sentences are constructed which follow the syntactic rules of a language but don't have a coherent meaning as a whole which makes it difficult to contextually predict the word. (Lemmetty, 1999), gives a good account of the evaluation tests for synthetic speech intelligibility.

It is difficult to evaluate the intelligibility of audio-visual speech. Synthesized AV speech is often tested for its most cited advantage over acoustic-only speech, i.e improvement in intelligibility in noisy conditions (LeGoff et al., 1994). Consequently, the addition of visual modality is evaluated by adding noise to the acoustic modality. This is because the intelligibility results of a visual-only speech would be very low, especially for SUS. On the contrary, in case of clear speech without any noise, the intelligibility us close to the best possible and does not add any additional advantage of visual modality. For instance, E.Cosatto et al. (2000) report that the AV speech shows significant improvement in terms of the intelligibility in noise when compared to acoustic speech, with an error rate of 4% for AV speech compared to 20% with acoustic speech. Fagel (2006) reports intelligibility tests of synthesized audio and AV speech in comparison with natural audio and AV speech. It was reported in terms of the percentage of vowel+consonant, vowel and consonant recognition errors. Ouni et al. (2007) present metrics to quantify the improvement in intelligibility between two visual conditions in comparison with acoustic-only speech.

In the methods which perform visual speech synthesis over acoustic speech, the synchronization of the two modalities is an additional aspect which needs to be evaluated. For example, Bregler et al. (1997) perceptually evaluate the lip-utterance synchronization, triphone-video synchronization i.e. the disruption level due to concatenation of units besides coarticulation effects. They report that there are occasional visible timing errors in the case of stop consonants and the visible articulation is unsatisfactory compared to the natural articulation of phoneme when the required phoneme sequence is not available in the corpus. Mattheyses et al. (2009) report a detailed perceptual evaluation of various image-based audio-visual speech synthesis techniques

to show the importance of audio-visual synchrony and coherence. The comparison was between the following 5 types of AV speech: (1) original AV speech, (2) AV speech synthesized by the concatenation of synchronous bimodal units, (3) AV speech synthesized by synthesizing audio and visual streams separately with the best audio and video segments respectively and then synchronizing them (4) visual and acoustic speech synthesis separately with their respective best segments, but the audio used for synthesis is from a different corpus, i.e., a different speaker to that of visual speech (5) AV speech with synthesized visual speech and real audio. The comparison was done to evaluate for audio-visual synchrony and perceived naturalness. The results of these perceptual comparative evaluation experiments favor audio-visual speech synthesis by synchronous bimodal-unit selection and concatenation. The results also show that the separate synthesis of the two modalities using different corpora is least preferable.

Sometimes a comparative evaluation of various systems is also done. Comparative evaluation of different approaches of speech synthesis is very useful. In the first place it provides a broad platform for the participants to evaluate their system performance. In addition, it brings out interesting directions to future research. Blizzard challenge started in 2005 by Black and Tokuda (2005) is one such platform. This annual challenge is designed for corpus based acoustic speech synthesis systems. The challenge provides a uniform framework to perform a comparative evaluation by removing the variability in database, test sentences being evaluated and the set of listeners evaluating the test sentences and finally the evaluation metrics. The set of listeners generally includes people from the following 3 categories: speech experts, volunteers and paid undergraduate students. The test sentences included sentences from 5 genres: novels, conversation, phonetically confusable sentences (Fairbanks, 1958; House et al., 1963) and semantically unpredictable sentences (Benoît et al., 1996). The initial 3 genres were for testing speech quality and the last two for testing the intelligibility of the synthesized speech. For quality evaluation, sentences synthesized by various synthesizers are played and listeners are asked to rank the quality in terms MOS score. Later on pairwise naturalness tests and speaker voice originality comparison tests were included. The latter test is more relevant for HMM based systems. The voice building has 3 variants from blizzard 2007 onwards, one using full corpus, the remaining 2 are based on using a subset of the speech corpus (Fraser and King, 2007). From 2008 blizzard challenge, the corpus had expressive speech also (Karaiskos et al., 2008). Later Blizzard challenges included evaluations of speech (1) for specific applications like telecommunications, human-computer interaction etc (King and Karaiskos, 2009); (2) in the presence of noise (King and Karaiskos, 2010); (3) intelligibility of names and addresses (King and Karaiskos, 2011). The notable analysis results of these evaluations are, that speaker voice originality and

join discontinuity has an affect on the quality evaluation, but intelligibility rather depends on join discontinuity alone (Clark et al., 2007).

LIPS challenge was a similar platform for evaluating visual speech synthesis techniques (Theobald et al., 2008). It was conducted for two years, 2008 and 2009. The aim is to eliminate the variability in the training data and evaluation related components like human subjects, test utterances and evaluation metrics. The training data was a one hour audio-visual corpus of a single speaker. The included utterances were phonetically balanced sentences, spoken in neutral speaking style without any expressions. The visual speech recording was in the frontal view such that all the articulators are clearly visible. The test utterances were 50 SUS sentences recorded in the same way as the training data (Benoît et al., 1996). The test utterances were provided as acoustic speech and hand-corrected phonetic transcript aligned with audio. Viewers were chosen from the INTERSPEECH-2008 conference participants with normal vision and hearing capabilities and who are English speakers. Synthesis systems were ranked for naturalness and intelligibility separately. For intelligibility acoustic component was degraded to signal-to-noise-ratio (SNR) of -10dB. Intelligibility was measured using speech recognition metrics defined in terms of insertions, substitutions and deletions. This was done by the comparison of identified and actual actual phonetic transcript. Visual speech naturalness was evaluated by asking the subjects to rate the synchronous audio-visual speech on a 5-point MOS scale. Such platforms provide communication grounds where the advantages and drawbacks of different approaches can be analyzed. This can pave way for the evolution of better techniques for speech synthesis.

2.6 Conclusion

We presented some aspects of unit-selection based speech synthesis. We have briefly discussed segmentation, and selection criteria for unit selection which included target cost and concatenation cost functions. We have also reviewed the general methodologies used to evaluate synthesized speech which are broadly divided into objective evaluation automatic evaluation and user-centered evaluation. The usage of a corpus does make it inflexible and might need effort to bring in changes due to the need to acquire and process a new corpus. Nevertheless, for any given application domain with specific requirements, it is always possible to build a unit-selection based speech synthesizer whose performance is comparable to real speech (Black, 2002).

Chapter 3

Acoustic-Visual Speech Synthesis System: An Overview

In this chapter we present an overview of our bimodal speech synthesis system named ViSAC. We refer to our system as acoustic-visual speech synthesis system to differentiate it from other classical approaches synthesizing acoustic and visual modalities separately. For us, speech is bimodal and the two modalities are kept together. We take this as the fundamental basis to our bimodal speech synthesis. Firstly, we record synchronous bimodal speech signal and process it to prepare the database. In this whole process, we keep the association of the two modalities intact. This results in a synchronous bimodal corpus. This database is then used by ViSAC to perform a concurrent synthesis of bimodal speech through unit selection. This proposed method implicitly addresses the problems of asynchrony and incoherence inherent in earlier classic approaches. The synthesis unit used by our system is diphone. The 3D data of the face is acquired during speech production using a stereo-vision technique simultaneously along with acoustic speech signal. The central synthesis paradigm is unit selection of bimodal segments. In audio-visual speech synthesis, required characteristics of both modilities need to be taken into account simultaneously. Hence, compared to acoustic-only speech synthesis, the problem complexity increases.

This chapter is organized as follows. We first detail the corpus acquisition and database preparation to be used for synthesis. Then, we describe the bimodal unit selection framework for acoustic-visual speech synthesis.

3.1 Corpus preparation

Unit selection is a corpus based synthesis methodology. The first step of corpus preparation involves careful text selection or design. It is done in such a way that the phoneme occurrence in the corpus is representative of the phoneme occurrence in the target language in general. Moreover, an effort is made to ensure at least minimum occurances of most of the synthesis units and good variants of the most frequent units. The uttered speech of the carefully chosen text is then recorded. The result of this is a speech realization for the underlying phoneme sequence specified by the text. The recorded speech is generally pre-processed for noise reduction when necessary. It is subsequently parametrized and segmented into phonetic segments. The text for which speech is recorded is not only analyzed in terms of its phonetic sequence but also for its detailed linguistic structure. In other words, we are interested in deriving any feature description from the text which can account for a variance in speech realization. We will call them target features. Thus, the phonetically segmented speech is annotated in terms of these target features extracted through text analysis. These segmented and annotated units constitute the speech corpus. To summarize, corpus preparation consists of the following stages:

- Text selection
- Data acquisition.
- Data processing and parameterization.
- Segmentation.
- Phonetic and linguistic annotation of the segmented data.

The final result of this corpus preparation in our case is a bimodal speech database. In the following subsections, we detail each of these steps, as performed for the preparing our audio-visual speech corpus.

3.1.1 Text selection

We built a corpus of a total of 319 sentences were recorded for the training corpus. It represents a total of 14634 diphones and includes a good variety of the most frequent diphones. Of course, this corpus doesn't cover a big variety of diphones, but our purpose is to experiment out methods. A set of 20 extra sentences were also recorded and set aside as the test sentences for evaluation purpose.

3.1.2 Acquisition

Visual data acquisition for our acoustic-visual speech synthesizer was performed simultaneously with acoustic data recording. It was done using a low-cost 3D facial data acquisition infrastructure developed by the team MAGRIT in our laboratory in the past (Wrobel-Dautcourt et al., 2005). The acquisition system uses two synchronized fast monochrome cameras (JAI TM-6740), a PC. During acquisition, the speaker with markers painted on his face, sat in front of a stereo camera pair with a microphone placed at 50-60 cm from his mouth. This whole technique provides a fast acquisition rate to enable an efficient temporal tracking of 3D points without affecting the speech articulation. Large majority of markers are detected by a low-level processing of the stereo image pairs (see fig. 3.1). This is based on their average gray-scale, shape and size (white circular points with a radius less than 3 pixels). Besides these points which are easily and accurately detectable, there are the following two cases:

- When the points cannot be detected directly. This occurs when some points are not visible in one or both of the images of some stereo image pairs. This might happen when the location of 3D marker is completely occluded during articulation like in the case of inner markers of lips. This can also happen when markers cannot be captured in one of the views due to the change in the head orientation.

- Where the detected points are not actual 3D markers. This is due to locations in the image with the same photometric features, as light reflects on eyes or teeth.

After the initial Processing, 86% of the 3D points are accurately reconstructed, 10% of the points are erroneous and 4% are missing which correspond to the hidden markers. Besides, the detection and reconstruction of marker, the markers are indexed for the creation of temporal trajectories based on temporal closeness. This indexing of the detected markers so as to indicate the location of the 3D marker on the face has occasional ambiguity. It happens mostly for markers on the lips, especially when they open and close. The markers which cannot be detected directly, they are estimated using an interpolation scheme that involves an initial 3D mesh of the face. This initial mesh is accurately built by automatic detection of 3D markers and subsequent correction was done by hand. Through this about 7% of the marker data is estimated in average. Processing the data is a lengthy work though. It takes several weeks for 28 minutes of data. The acquisition of the bimodal corpus, the stereoimage processing and 3D marker extraction was done by members of team MAGRIT, who are a part of this project.

The recorded corpus consisted of the 3D positions of 252 markers covering the whole face. However, the lower face was covered by 70% of all the markers (178 markers), where 52 markers

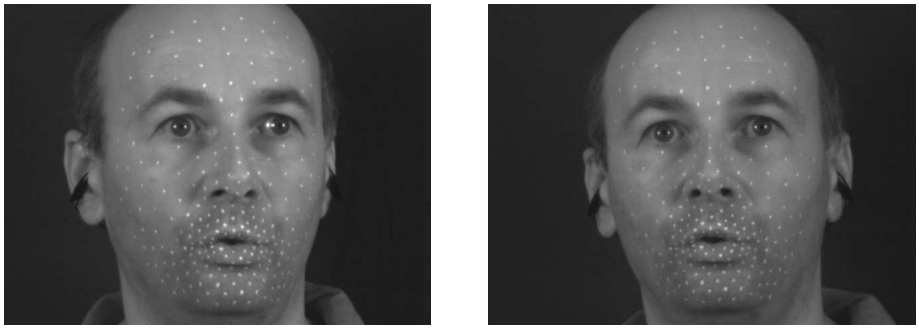

Figure 3.1: *Stereo-vision image pair of the speaker*

were covering only the lips. This choice was made, to capture the lip movement accurately and to be able to model the lips finely. The average sampling rate was 188 Hz. The corpus was made of 319 medium-sized French sentences, covering about 28 minutes of speech, uttered by a native male speaker. An extra 20 sentences were recorded for testing purposes. The speech signal was recorded at 16 kHz with 16-bit precision.

3.1.3 Data processing and parameter extraction

The sampling rate of the acquired 3D marker data was around 188Hz. There was a slight variance in the sampling rate across sentences. A set of sentences were recorded in different sessions with short pauses between successive sessions. This variance in the acquired data is due to a slight variable lag between the time instant the images were captured and sent to the computer for storage. The data was filtered using a low-pass filter with a cut-off frequency of 25 Hz. Such a processing removes additive noise from the visual trajectories without suppressing important positional information.

Principal Component Analysis (PCA) was applied on a subset of markers of the lower part of the face (jaw, lips, and cheeks; see Fig 3.2). The reason for this choice was that the movements of markers on the lower part of the face are tightly connected to speech gestures. Markers on the upper part of the face either do not move, or their movements are of no direct relevance to speech. This can be said because the speech is recorded with a neutral voice with no strong prosodic effects. We have not used any guided PCA as it does not provide significant advantage. Besides,the projection onto principal components and reconstruction are straightforward and fast. This unified approach keeps it simple and straight forward for the synthesis purpose. The facial deformations when each of the principal components is set at -3 and 3 z-scores is shown in figure 3.3. The first two components account for 79.6% of facial speech data variance. It is difficult to draw definite conclusions about the influence of each principal component on facial deformation. The affect of each of the principal components cannot be completely isolated in

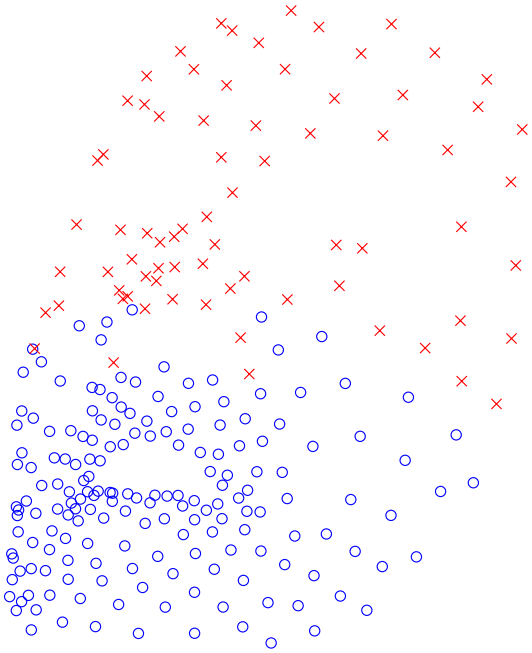

Figure 3.2: *PCA is applied on 178 (plotted as blue circles) out of 252 painted markers.*

terms of the perceived facial deformations. Broadly, the following observation can be made by looking at visual speech animation by varying a single principal component. The first two principal components mainly account for combined jaw opening/closing and lip protrusion gestures. The third component accounts for lip opening, after removal of the jaw contribution. Some of the components though related to speech, are augmented by some gestures that are specific to speaker's facial expressions. This seems to be the case for components 4 and 5. They seem to capture lip spreading. However, due to some asymmetry in our speaker's articulation, lip spreading is divided into two modes: one accounting for spreading toward the left side of the lips and one for spreading toward the right side. Component 6 is a smiling gesture, however it is difficult to classify it as belonging to speech articulation or pure facial expression. Components 7 to 12 seem to account for very subtle lip deformations, which we believe are idiosyncratic characteristics of our speaker.

Several experiments indicated that retaining as less as three components could lead to an animation which would be acceptable, in the sense that it would capture the basic speech gestures and would filter out almost all the speaker specific gestures. However, such an animation would lack some naturalness, which is mostly captured by secondary components. We are also in favor of keeping the specificity of the speaker specific gestures. Retaining 12 components leads to animations that are natural enough for all purposes. One of the goals of our proposed system is to synthesize trajectories corresponding to the PCA-reduced visual information, for these 12 components, alongside the synthesized acoustic speech signal. The lower face related visual

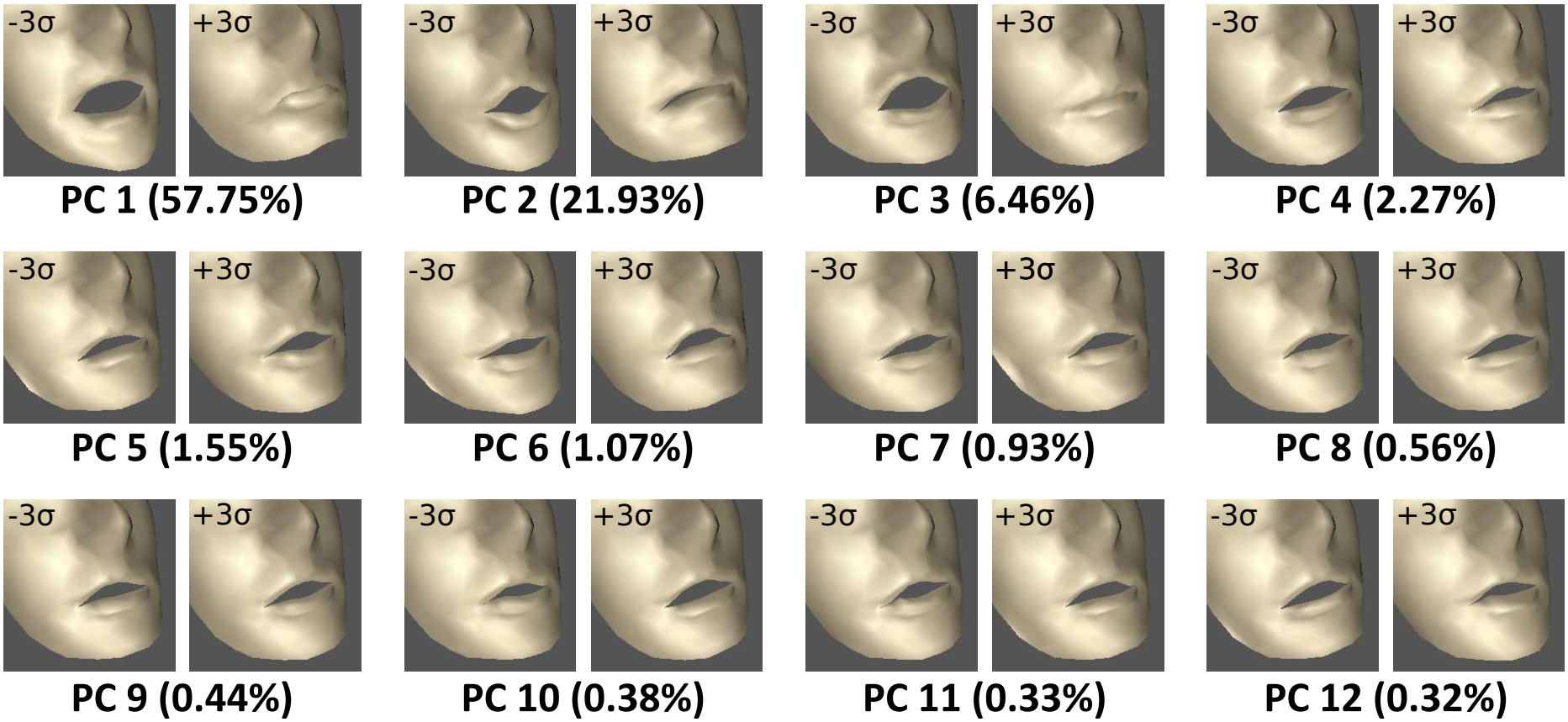

Figure 3.3: *Facial deformations when each of the principal components is set at* -3 *and* 3 *z-scores.*

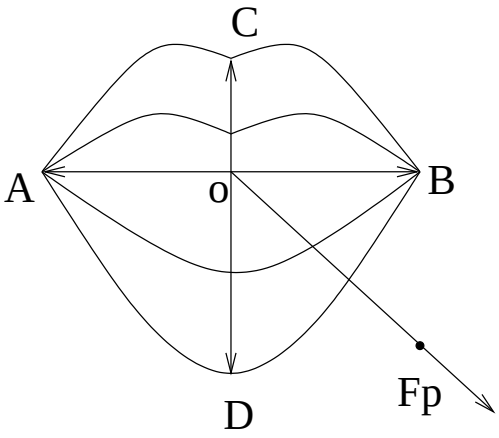

Figure 3.4: *Calculation of labial features is done using the 4 points on the face: A, B, C and D. Lip opening and lip spread are given by the distances $\|\vec{CD}\|$ and $\|\vec{AB}\|$. Lip protrusion is given by the displacement of O, the center of gravity of the four points (A, B, C, D) along the normal vector ($\vec{OFp}$) to the plane formed by vectors $\vec{AB}$ and $\vec{CD}$. Jaw opening is calculated as the distance between the center of the chin and a fixed point on the head.*

information can be reconstructed using these 12 trajectories. The mean values of the positions of the markers at the upper part of the face may then be added to complete the face visualization. Hence, the 12 first principal components, which explains about 94% of the variance of the lower part of the face are retained for storage and reconstruction at runtime. Besides the 12 PCA coefficients, four articulatory parameters (lip protrusion, lip opening, lip spread and jaw opening) are calculated as explained in figure 3.4) (Robert et al., 2005). These articulatory features are used for the analysis of visual speech corpus and during implicitly during selection as visual target costs are designed based on these features.

The acoustic speech paramters extracted included the LPC (Linear predictive coding) coefficients, f0, and energy.

3.1.4 Segmentation

We perform segmentation based on the forced alignment of acoustic speech. These predicted segment boundaries are considered as the synchronous bimodal segment boundaries, and chosen to represent speech segments in the corpus. The synthesis unit of target search and synthesis is the diphone. Besides making the storage and indexing of bimodal speech segments extremely simple, it reinforces the principal idea of synchronous inseparable bimodal speech intact. A diphone extends from the mid of one phone to the mid of the next phone. The middle of the phone is a relatively stationary region. Hence by using diphone as the synthesis unit, the acoustic artifacts due to any segmentation errors are reduced. Diphone units also account for the coarticulation well, as their boundaries include the transition of one phoneme into the other.

Diphone as a synthesis unit is reported to produce comparatively good quality speech (Moulines and Charpentier, 1990). The Segmentation based on speech acoustics and annotation of data was done using scripts developed by Colotte (2009). The monophone HMMs which are used by these scripts are trained on a very large acoustic speech corpus and provide highly accurate segmentation.

3.1.5 Bimodal speech database

The phonetized corpus was analyzed linguistically, and partitioned into phonemes. To mark the diphones from these phonemes and describe them in terms of target features, we used tools that have been already developed in the framework of SoJA Colotte (2009). For each phonetic unit in the corpus, the following information is included for its indexing:

- The description in terms of the complete target feature set (Fig. 3.6).
- Its position (start sample to end sample) in the corresponding acoustic and visual speech data files.
- Duration.
- Acoustic and visual parametric representation at the middle of the phonemes that we have extracted (section 3.1.3).

The phonetic and linguistic annotation of the speech units is taken from SoJA.

3.2 Bimodal speech synthesis

Our Text-to-Speech (TTS) Synthesis system has two stages. First stage is the Natural Language Processing (NLP) stage which analyzes the input text. It provides as a result, the specification of the target phoneme sequence required for synthesis. This specification is represented using a combination of target features based on the linguistic and phonetic structure of the text. The second stage involves the actual speech synthesis for the required target sequence using bimodal unit selection and concatenation.

3.2.1 Natural language processing

The first stage of our TTS system is an NLP unit. For a given text, it generates the phoneme sequence from text to be synthesized. As shown in fig. 3.5, this is done by following these steps (see fig. 3.5):

- **Preprocessing:**
  - ◇ *Text Segmentation:* Input text is split into individual sentences which can be processed separately.
  - ◇ *Tokenization:* Each sentence is split into tokens depending on breaks based on white spaces, punctuation marks etc. It is done so that they can be analyzed separately. Each token is classified into different classes such as words, numbers, dates, abbreviations etc. This is done to determine the kind of parsing and verbalization to be done if necessary.
  - ◇ *Parsing:* Each non-natural language token is parsed to decode the exact format of the text.
  - ◇ *verbalization:* Each decoded/parsed non-natural language token is verbalized into words.
- **Lemmettization:** Each of the tokens is morphologically analyzed, and all the probable root forms of the words are enlisted.
- **Tagging:** Each of the tokens is then syntactically tagged with the most probable part of speech pin-pointing the word in the dictionary .
- **Chunking:** The phoneme sequence is divided into rhythm groups using chunker based on some rules. This is similar to phrasing done for English.
- **Phonetization:** Words are phonetized into phoneme sequences after homograph disambiguation wherever necessary. This is done using lexicons. There are different lexicons based on the kind of words. Words are classified into different groups like French word, proper noun, word belonging to a foreign language etc. Depending on this category, the appropriate dictionary is used to give the word to phoneme sequence mappings. For words which are not present in any of the lexicons being used by the system, listed grapheme-to-phoneme rules are applied.
- **Post lexical processing or post phonetization:** In languages such as French, the words interact with each other to produce different phoneme sequences based on some specific rules. Hence, the phonetized text is re-analyzed for continuous-speech related rules like liaison to modify the phoneme sequence.
- **Syllabification:** The phoneme sequence is divided into syllables based on rules. Rhythm groups and syllables, these two units are known to be important for explaining various

aspects of prosody for French.

3.2.2 Target unit description

Each phoneme in the text is described in terms of linguistic and phonetic features which are known to affect the acoustic and visual realization of the phoneme. The target (resp. candidate) specification (resp. description) is done in terms of their characteristics at various levels as shown in figure 3.6.

3.2.3 Bimodal unit selection and concatenation

The target sequence is based on phonemes, that are specified after the text analysis and converted into diphone-based targets. For each required target diphone, all possible candidates from the corpus which have the same phonemic label are looked up. The specification of targets for synthesis is in terms of the same features used to describe the candidates in the corpus. These descriptive features are exhaustive phonetic and linguistic features that can be extracted. They can be either independent or dependent on the target language. This target specification is compared with that of the description of the candidates in the corpus. For a target sequence specification $t_1^n = (t_1, ...t_j, ...t_n)$, a general target cost function TC is calculated as follows:

$$TC = C(t_i, u_i) = \sum_{\rho=1}^{F} w_\rho C_\rho(t_i, u_i) \tag{3.1}$$

where, $C_\rho(t_i, u_i)(\rho = 1, ..., F)$, are the different target feature costs between a target t_i and a candidate u_i, F is the total number of target features and w_ρ is the weight given to a feature ρ. The acoustic join cost is defined as the acoustic distance between the units to be concatenated. It is calculated using the following acoustic features at the boundaries of the units to be concatenated:

- fundamental frequency (f0).
- LPC coefficients.
- Energy.
- Duration.

Similarly, the visual join cost is defined as the visual distance between the units to be concatenated as shown in figure 3.7. This is calculated using the PCA transformed visual

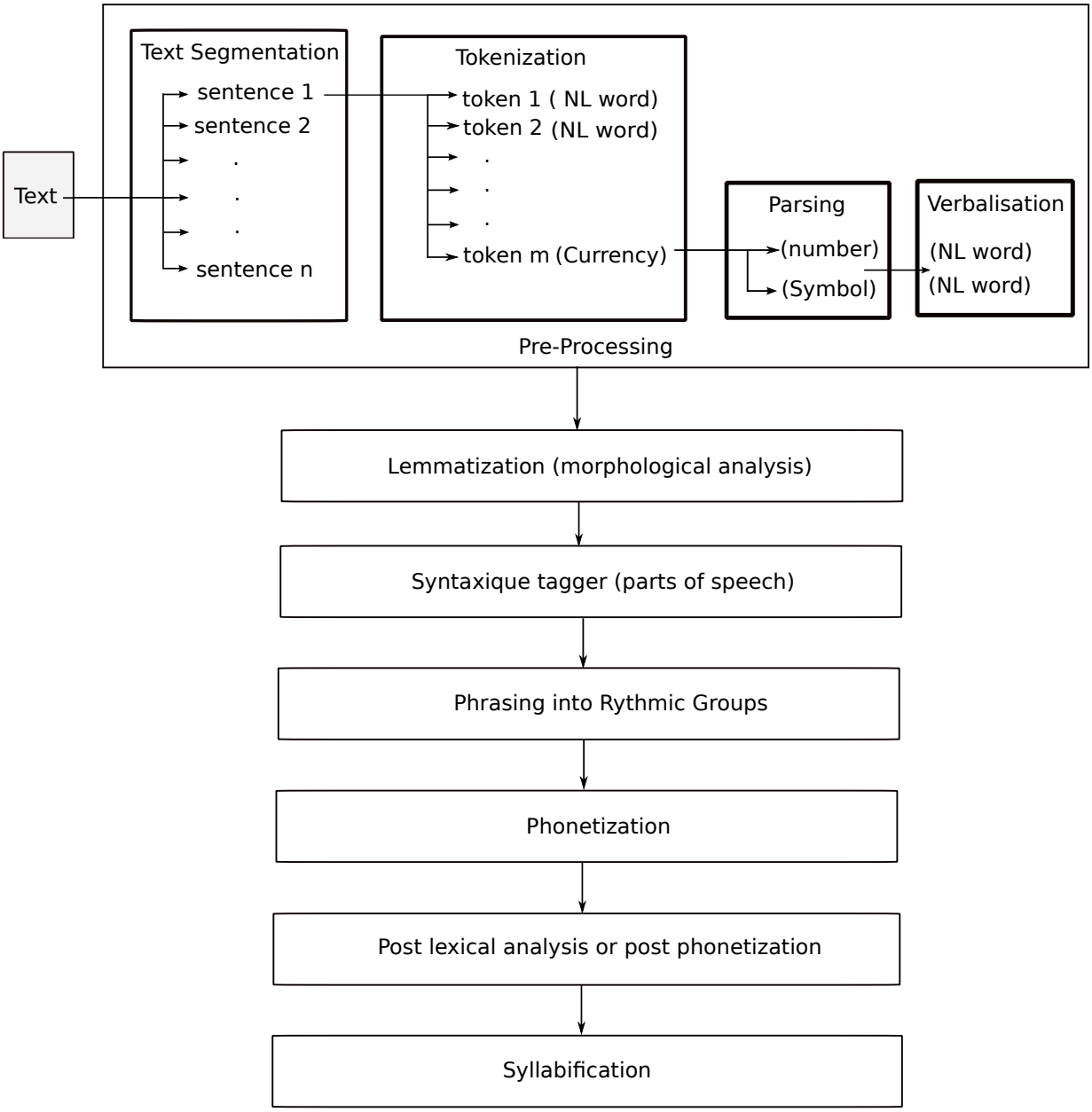

Figure 3.5: *Text processing to output the necessary target phoneme sequence to be synthesized*

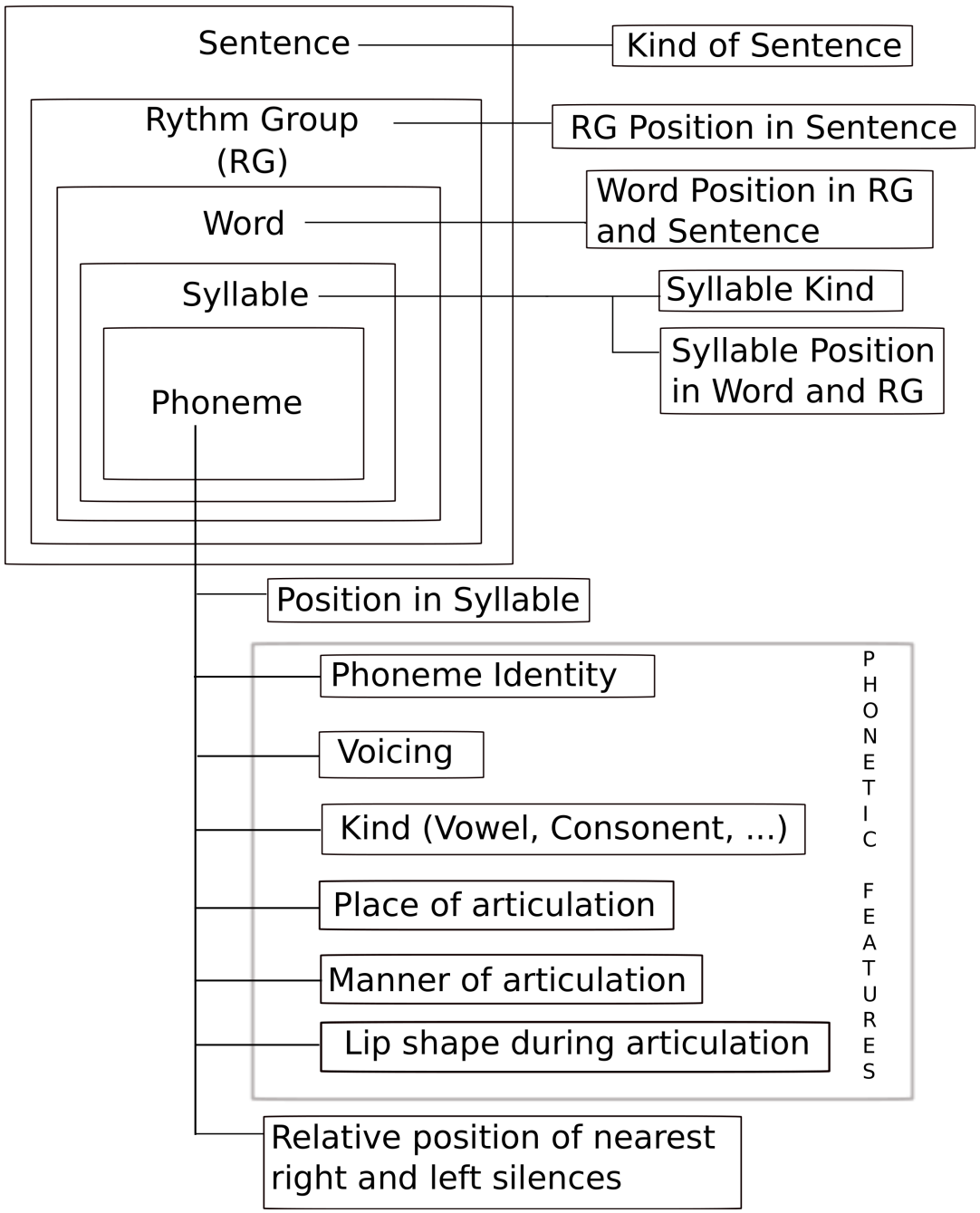

Figure 3.6: *Target phoneme specification using phonetic and linguistic descriptors at various levels*

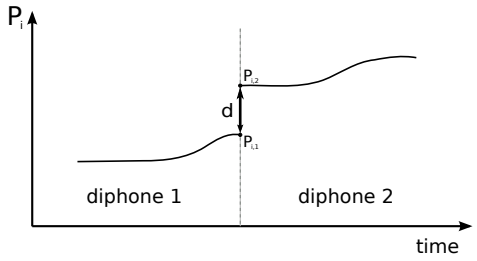

Figure 3.7: *An illustration of the gap in the visual feature trajectories. The purpose of the visual join cost is to minimize the discontinuities in the visual modality at the boundaries where concatenation happens.*

.

information at the boundaries of the units to be concatenated. That is:

$$VC = \Sigma_{i=1}^{12} w_i (P_{i,1} - P_{i,2})^2$$

where $P_{i,1}$ and $P_{i,2}$ are the values of the projection on principal component i at the boundary between the two diphones. The choice of weights w_i is based on the relative importance of the components. We chose these weights to be proportional to the eigenvalues of PCA analysis as they are proportional to the data variance accounted by the respective principal component. This is similar to the methodology mentioned in (Liu and Ostermann, 2009). The selected diphone sequence is concatenated acoustically using a traditional technique, where pitch values are used to improve the join of diphones.

The selection among the set of pre-selected candidates is operated by resolving the lattice of possibilities using the Viterbi algorithm. The result of the selection is the path in the lattice of candidates which minimizes a weighted linear combination of the following three costs:

- Target cost ($\sum_{i=1}^{n} C(t_i, u_i)$).
- Acoustic join cost ($\sum_{i=2}^{n} C^{aj}(u_{i-1}, u_i)$).
- visual join cost ($\sum_{i=2}^{n} C^{vj}(u_{i-1}, u_i)$).

It is calculated as follows:

$$C^T(t_1^n, u_1^n) = \min_{u_1,\ldots,u_n} \begin{cases} w \sum_{i=1}^{n} C(t_i, u_i) & + \\ w_{aj} \sum_{i=2}^{n} C^{aj}(u_{i-1}, u_i) & + \\ w_{vj} \sum_{i=2}^{n} C^{vj}(u_{i-1}, u_i) & \end{cases} \tag{3.2}$$

where w, w_{aj} and w_{vj} are weights for the component target cost, acoustic join cost and visual join cost, the weights used are $w = 1$, $w_{aj} = 0.943$ and $w_{vj} = 0.897$ (Toutios et al., 2011). I

have participated in developing the first version of ViSAC, but it was mainly developed by A. TOUTIOS in collaboration with V. Colotte and S. OUNI. A synthesis example of one of the test sentences is given in figure 3.8.

3.3 Visual speech rendering

The visual speech in ViSAC is rendered as a face approximated using sparse 3D mesh, but two alternatives are also included. We didn't add a tongue yet. This appearance of the 3D-marker rendering, wired mesh surface made with the 3D-marker data and the face approximated using the sparse meshes are shown in figure 3.9. A simple visual speech animation of the syllable 'ba' is shown in the figure 3.10.

3.4 Conclusion

In this chapter, we described corpus acquisition and database preparation for our system. We presented an overview of our text to acoustic-visual speech synthesis system called ViSAC. The synthesized speech with this initial system clearly indicated the advantage of synchronous bimodal unit concatenation. Besides, this framework presented the experimental setup for developing various methodologies for improving bimodal speech[1].

[1]Parts of the work presented in this chapter was published in (Toutios et al., 2010a) and (Toutios et al., 2010b).

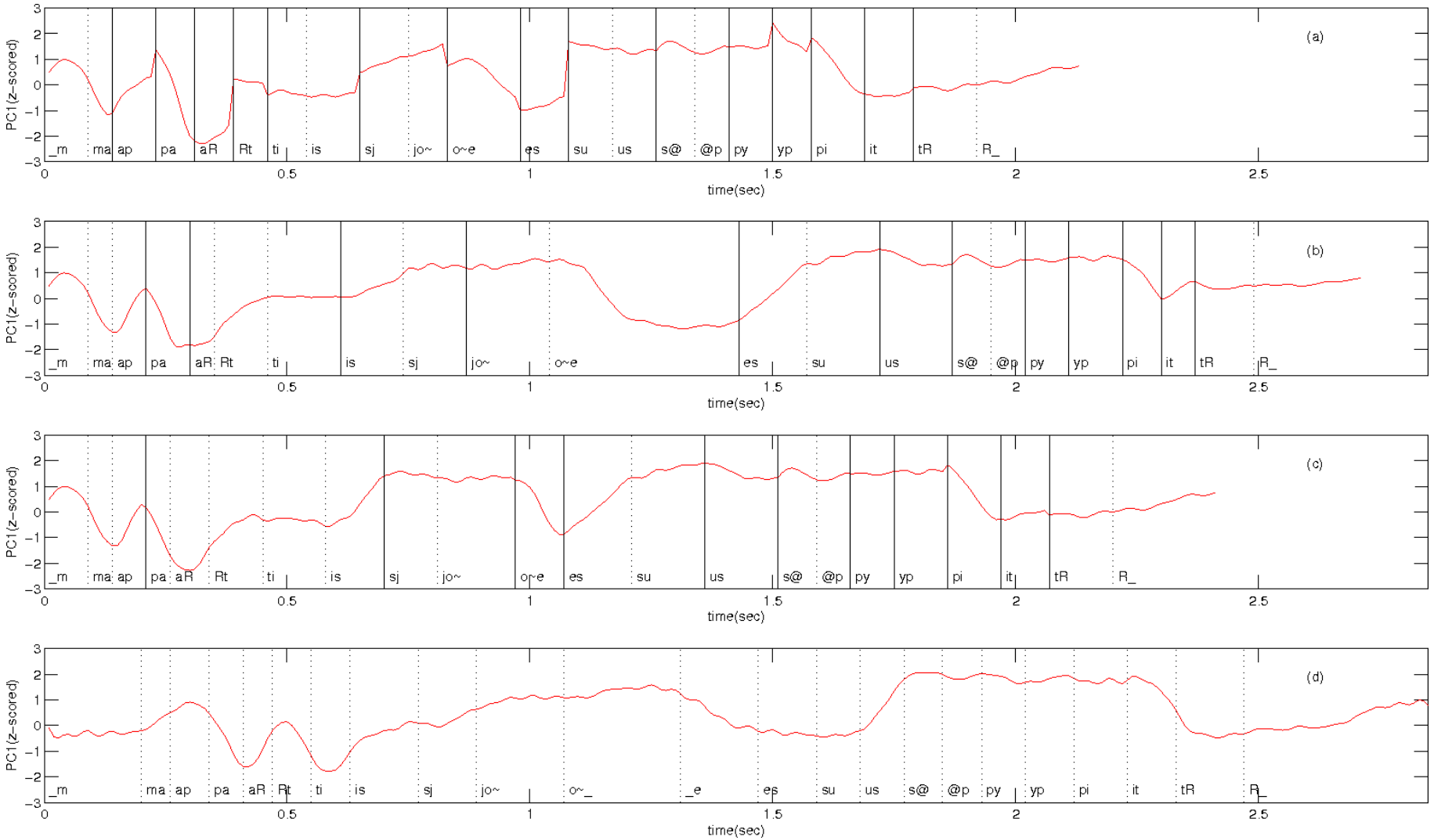

Figure 3.8: *A synthesis example showing the trajectory of the first principal component. Figures (a), (b) and (c) show the trajectories synthesized with acoustic-only, visual-only and audio-visual join costs. Figure (d) gives the first principal component of the real sentence*

.

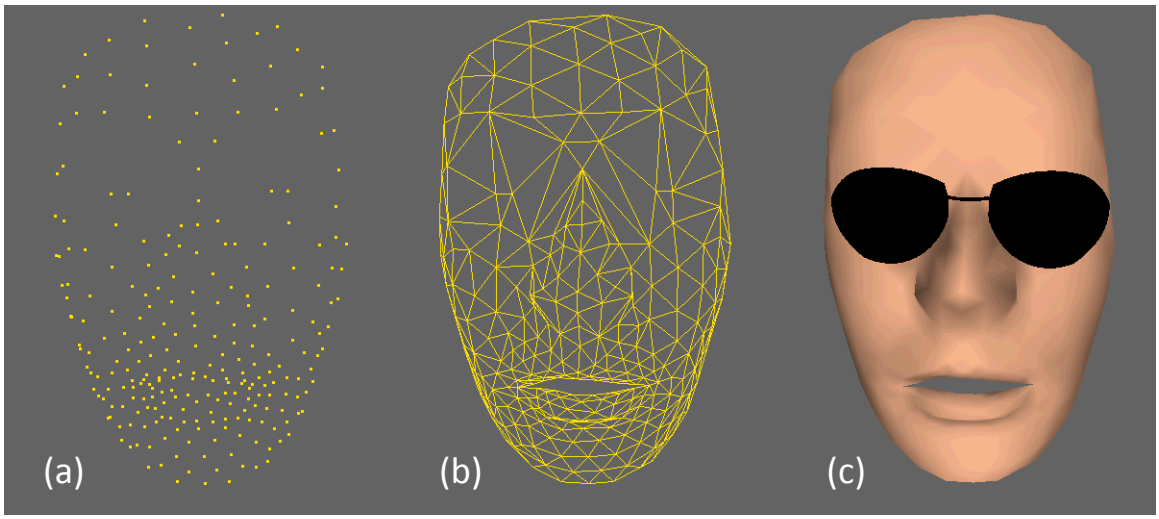

Figure 3.9: Shows the appearance of (a) just the 3D-marker rendering, (b) wired mesh surface made with the 3D-marker data and (c) the face approximated using the sparse meshes.

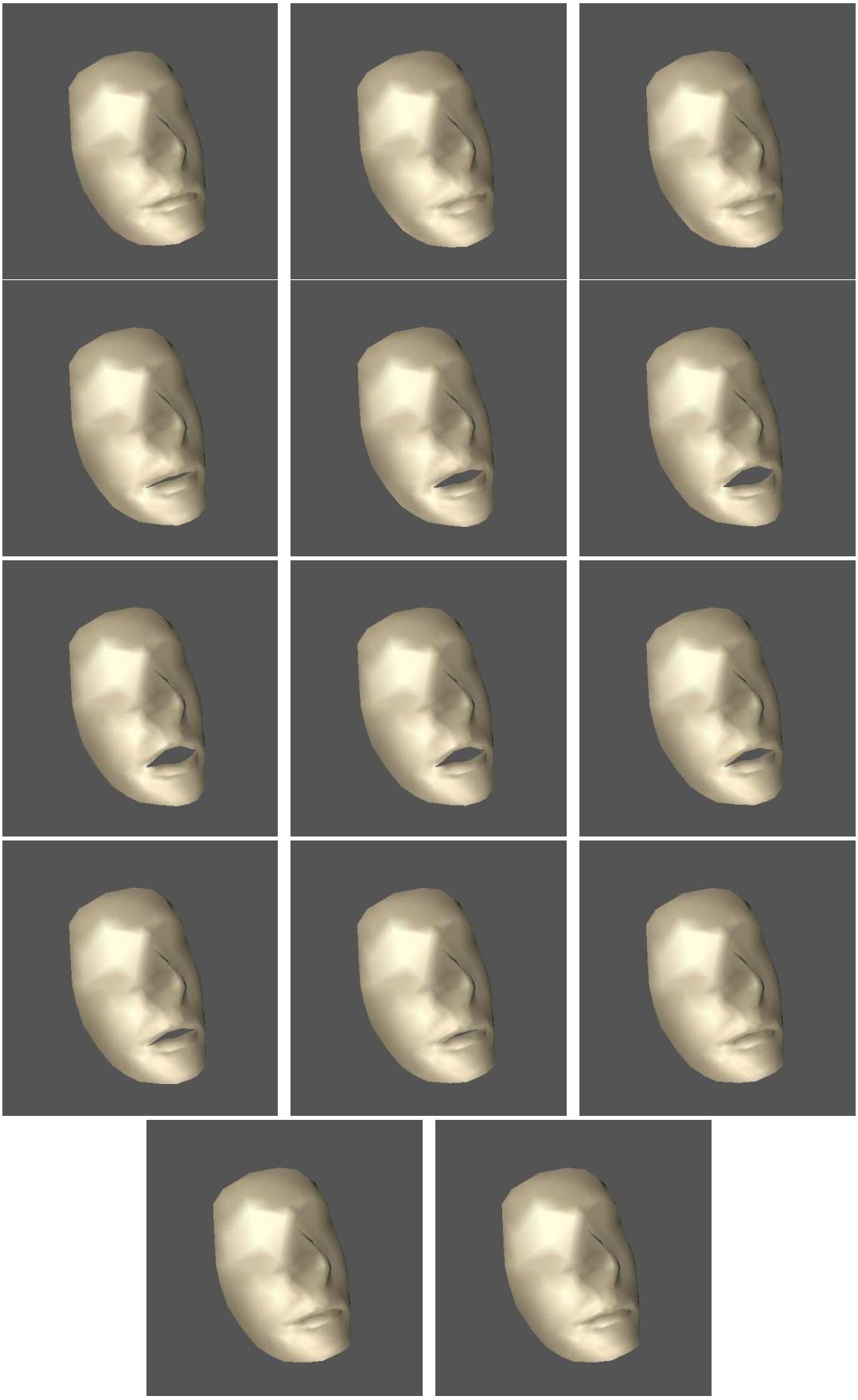

Figure 3.10: Visual speech of the syllable "*sil* b a *sil*" with a frame rate of 25fps.

Chapter 4

Phoneme Classification Based on Facial Data

The facial data that we have acquired, only encodes the speech related deformations of the outer surface of the face. This kind of data lacks internal articulatory information. It would be interesting to investigate the representative phonetic patterns in this kind of data. It might also give an estimate of the articulatory information that is lacking in this kind of data. Keeping these objectives into account, we have performed some segmentation experiments. First we used our facial data. Then, to estimate the internal articulatory information that is missing in comparison to the facial data, we performed another set of segmentation experiments. This time, we used a different corpus which had articulatory information related to the tongue.

In the following sections, we describe these two sets of experiments, first using our facial data in section 4.1 and then using an EMA (Electromagnetic articulography) data.

4.1 Visual speech segmentation using facial data

Phonetic boundaries are generally used to segment bimodal speech corpus. Though this is the case, the start and end of the acoustic speech and visual speech gestures might happen at different time instances. This is because, for the sound production to happen, the prior articulatory configuration required for the production of sound has to be attained first. The time differences between acoustic and visual segment boundaries might probably vary due to coarticulation. Phonetic units which are segmented using acoustics thus might not capture the start and end of the segments in the visual modality accurately. But, these acoustic boundaries would give an indication of approximate time intervals of the phoneme articulation. Ideally, segmentation based on visual speech should provide us this information. By following this

rational, an elaborate experiment was performed to segment the visual speech using the facial data. The contributions of this experimental results are two-fold. They provide significant information about the uniqueness of phonetic articulation accounted by just the facial data which might be perceived more accurately by humans. Due to this humans might also be more critical about the facial animation of such phonemes. They also provide information about which phonemes are influential or are influenced in the context of other phonemes.

In order to segment the visual speech data, we trained phoneme HMMs using a procedure similar to the one typically used in Automatic Speech Recognition (ASR). We used HTK for this purpose (Young et al., 2005). We used three different feature vectors extracted from the facial data. The three sets of feature vectors used for HMM training are the following:

- Articulatory features.
- PCA coefficients.
- Combination of the articulatory and PCA coefficients

The set of labels include the set of phonemes covered in the corpus and *sil* (silence). One monophone HMM is trained for each of the labels in this set. The HMM training performed is similar to that performed for a conventional ASR module. In the first step, monophone HMMs corresponding to each label were trained. Each HMM was a 3-state left-to-right no-skip model. The output distribution of each state was a single Gaussian with a diagonal covariance matrix. The observation vectors input to the HMM training consisted of static and dynamic parameters, i.e. the three types of feature vectors described in the previous section and their delta and delta-delta coefficients. The HMM parameter estimation was based on the ML (Maximum-Likelihood) criterion estimated using Baum-Welch recursion algorithm. The learned monophone HMMs were used to perform a forced alignment of the same training corpus.

Forced alignment was performed with three sets of monophone HMMs trained using the three feature vectors. The HMM training is an iterative process. To evaluate the segmentation, we have used a recognition criterion explained in the following subsection. For each set of HMMs trained using a particular set of feature vectors the following is done. After each iteration of HMM parameter re-estimation, the training data is segmented using the updated HMMs. Then, the total recognition error of the segmentation is calculated. Training is halted when there is no further improvement in this value in subsequent iterations. The recognition error of each labeled visual segment in the corpus at this stage has been used for the evaluation and analysis of the alignment results. The set of monophone HMMs which gave the best segmentation result based on the total recognition error was chosen for the second step for further improvement. The

second training step involved creation of context dependent triphone models using the trained monophone HMMs. Finally tied-state triphones were created using decision tree clustering. The triphone models were created by first cloning the trained monophone HMMs for different triphones. Then, triphones which have sufficient data in the corpus are re-estimated. Then using decision tree clustering, tied state triphones were created. The contexts considered for clustering are based on the hierarchical cluster trees of phonemes mentioned in (Odisio et al., 2004). The complete speech corpus has been used for the estimation of HMM parameters. These trained HMMs were then used to perform forced alignment of the data. An example of the segmentation through the HMMs which are trained using the facial data is shown in Figure 4.1.

4.1.1 Recognition error

It has been shown that visual speech segments are correlated to the corresponding acoustic speech (Barker and Berthommier, 1999; Yehia et al., 1998). In fact, the speech sound is the consequence of the vocal tract deformation and thus the face. Thus, there has to be an overlap between the actual acoustic and visual speech segments. The visual and acoustic speech segments might have asynchrony in their onset and end time as the vocal tract has to anticipate the following sound by adjusting the different articulators.

Based on the above reasoning of asynchrony and overlap of the visual and acoustic speech, we have derived the following criterion for evaluating the segmentation results. We consider the recognition of a label to be correct, if there is an overlap between the predicted visual segment and the actual acoustic segment, the overlap being however small. An ASR engine trained with a very large acoustic corpus was used to provide the phoneme labels and acoustic boundaries of our acoustic whole corpus. We consider the acoustic boundaries given by the ASR engine as the accurate acoustic boundaries for comparison.

4.1.2 Forced alignment results

In this subsection, we present the quantitative results based on the recognition error mentioned in the previous section. We classify phonemes based on their visibility as shown in Table 4.1. We consider /ɥ/, /w/, /ʃ/ and /ʒ/ as bilabial based on their secondary place of articulation. In fact, their primary place of articulation is not relevant to our study (not visible) as it is the case for the secondary place of articulation.

We performed 4 alignment experiments. These include 3 experiments based on training monophone HMMs using the 3 types of feature vectors mentioned above. Based on the alignment results with the 3 sets of monophone HMMs, the feature vector performing the best among the

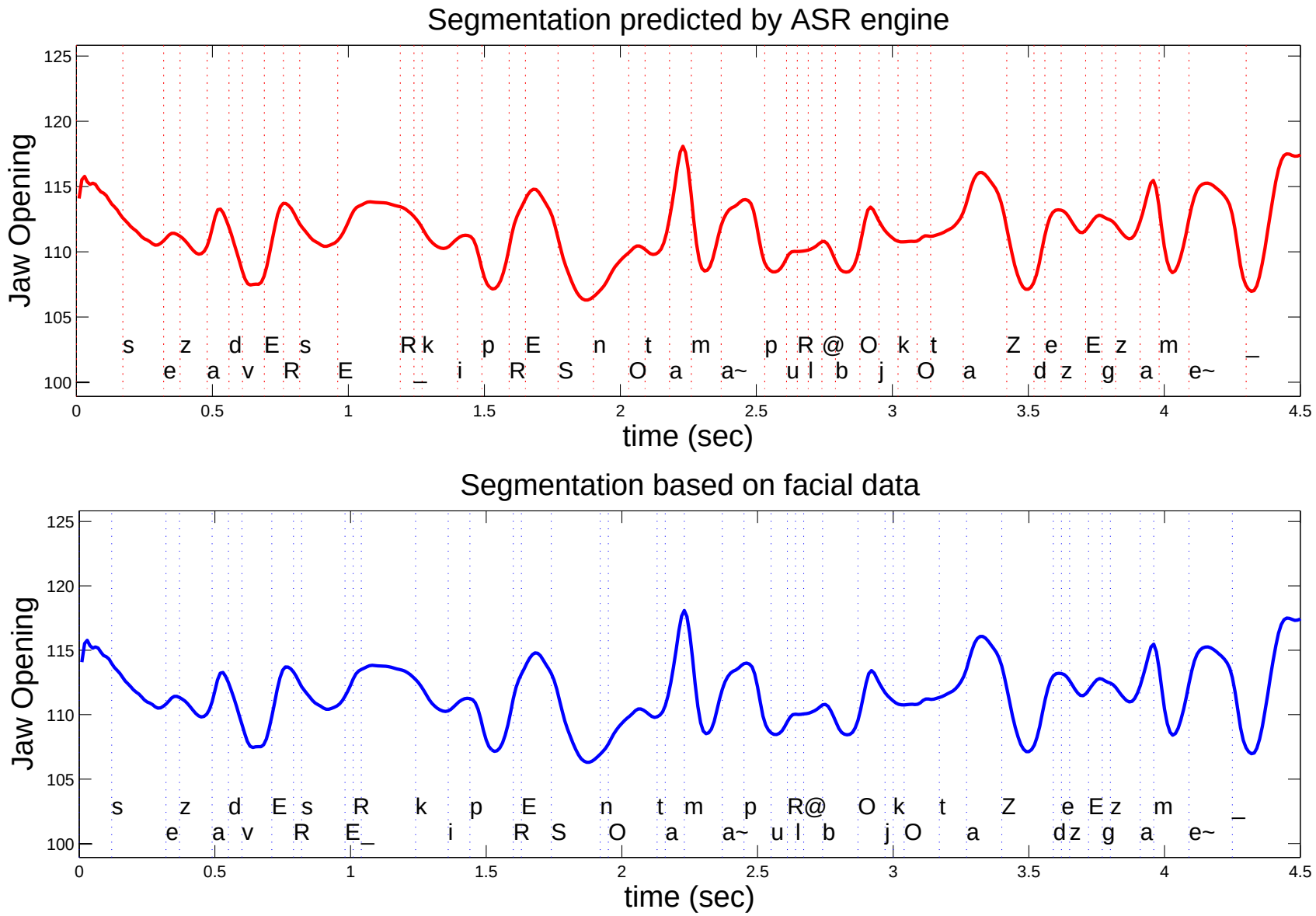

Figure 4.1: *An example of the segmentation using the HMMs trained with facial data. It is shown in comparison with the segmentation performed using ASR engine. The jaw opening is expressed in terms of relative units calculated based on the 3D coordinates.*

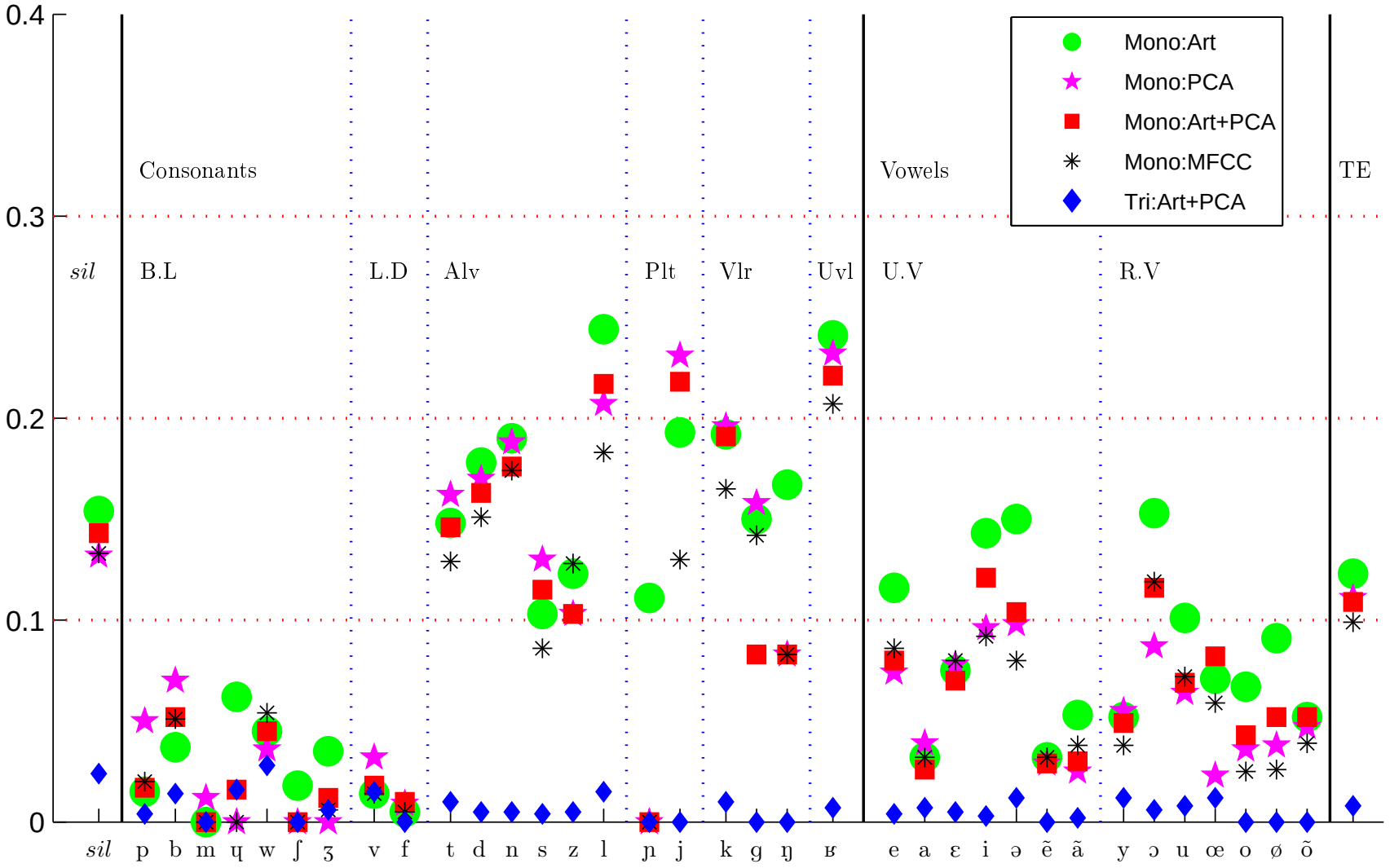

Figure 4.2: *Recognition errors in the alignments: Art, PCA are the articulatory and principal component analysis based feature vectors; MFCC are the acoustic parameters (Mel-frequency cepstral coefficients); Mono and Tri are the monophone and triphone HMMs. TE is the total recognition error. An automatically predicted segment alignment is considered correct when it has some overlap with the corresponding acoustic segment, however small it might be. Based on the low recognition error, the set of following phonemes can be classified as being visible: { p, b, m, ɥ, w, ʃ, ʒ, v, f, y,a, ã, o, õ, ẽ }*

Vis.	Abbr.	Class	Members of the class
1	B.L	bilabial	p, b, m, ɥ°, w°, ʃ°, ʒ°
	L.D	labiodental	v, f
	R.w	rounded vowels	y, ɔ, u, œ, o, ø, õ
2	*sil*	*sil*	*sil*
	Alv	alveolar and dental	t, d, n, s, z, l
	Plt	palatal	ɲ, j
	vlr	velar	k, g, ŋ
	Uvl	Uvular	ʁ
	U.V	unrounded vowels	e, a, ɛ, i, ə, ẽ, ã

Table 4.1: *Classification of phonemes based on their visibility. Phonemes classified as 1 are visible and 2 are invisible. Phonemes followed by ∘ are classified based on their secondary place of articulation.*

three based on the total recognition error (section 4.1.1) was selected for training the context dependent triphone models for further improvement of alignment. The results are presented in Figure 4.2. The PCA based feature vectors perform better than articulatory feature vectors in terms of the total recognition error. The heterogeneous feature vector, consisting of both PCA based features and articulatory features, performs better than each taken alone. PCA based features quantitatively account for the overall shape or deformation during the speech production. The articulatory parameters increase the discrimination by quantifying the typical articulatory characteristics like complete closure of mouth for /p/. This performance is further improved by triphone HMMs. As one can expect the recognition errors are low for phonemes which involve labial region for their coarticulation. The recognition errors are relatively higher for other consonant classes.

To verify that substantial training can be achieved by our small corpus (28 minutes of audio-visual speech), monophone HMMs were trained using the acoustic speech of our corpus. The acoustic features extracted from the speech were the MFCC (Mel-frequency cepstral coefficient) features vectors. The trained HMMs were used for the forced alignment of the same speech data that was used for training. The resulting acoustic segments were compared with the segments predicted by the ASR engine. The total recognition error used to quantify the visual segmentation results was determined in this case. A total recognition error of less than 1% was observed. Based on the low recognition error, looking at the figure 4.2, the set of following phonemes can be classified as being visible: { p, b, m, ɥ, w, ʃ, ʒ, v, f, y,a, ã, o, õ, ẽ }. These phonemes have a component of unique articulatory information embedded in the facial data. Thus, these phonemes need more importance in synthesis of visual speech animation using this kind of facial data.

The following analysis has been done considering only the correctly recognized visual seg-

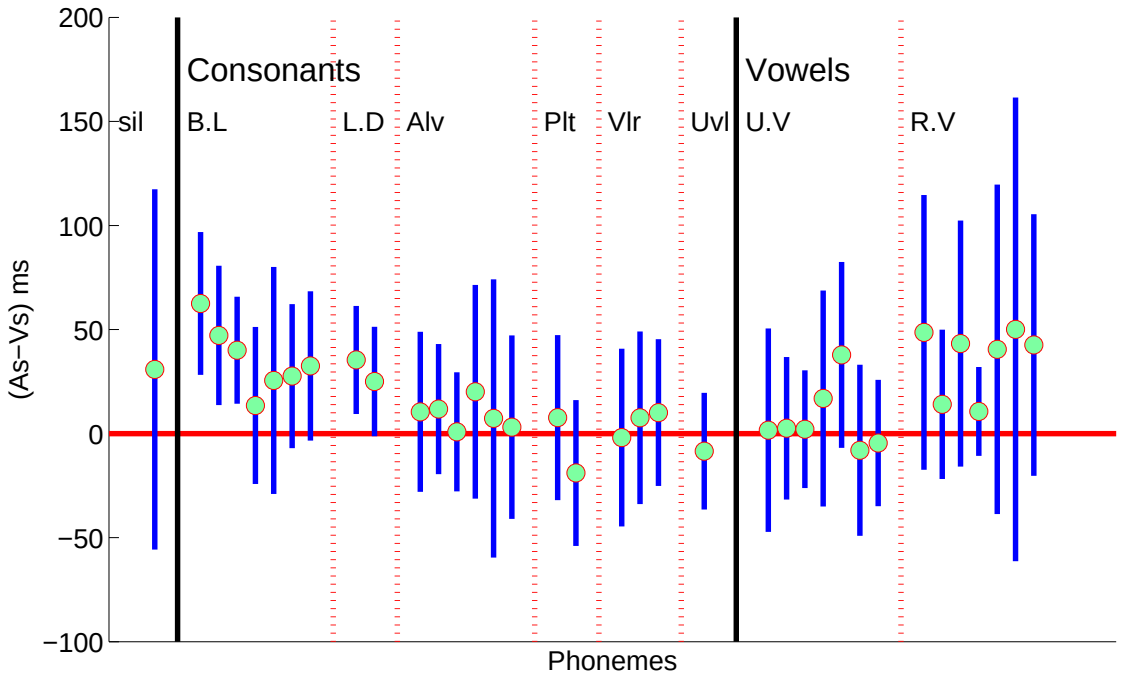

Figure 4.3: *Mean difference in the starts of acoustic and visual speech segments*

ments. Let As and Vs be the starts of the acoustic and visual segments of the same phonetic label, Ae and Ve be the ends of the acoustic and visual segments of the label. Let Ds be the start difference and De be the end difference, calculated as follows:

$$Ds = (As - Vs),$$

$$De = (Ae - Ve)$$

The mean and variance of Ds and De are calculated for each of the labels covered by the corpus (see Fig. 4.3 and Fig. 4.4). In the following analysis, focus has been given to only those phonemes which have significant coverage in the corpus. A positive expectation of the start difference, $(E(Ds) > 0)$ means visual start leads over the acoustic start. This suggests a visual influence of the speech coarticulation on the left contexts. This is the case for bilabials, labiodental and rounded vowels. Similarly, $(E(De) < 0)$ means acoustic end leads over visual end, with a visual influence of the speech coarticulation on the right context. The segmentation results that was obtained show that /ɥ/, /w/, /ʃ/ and /ʒ/ fall in this category.

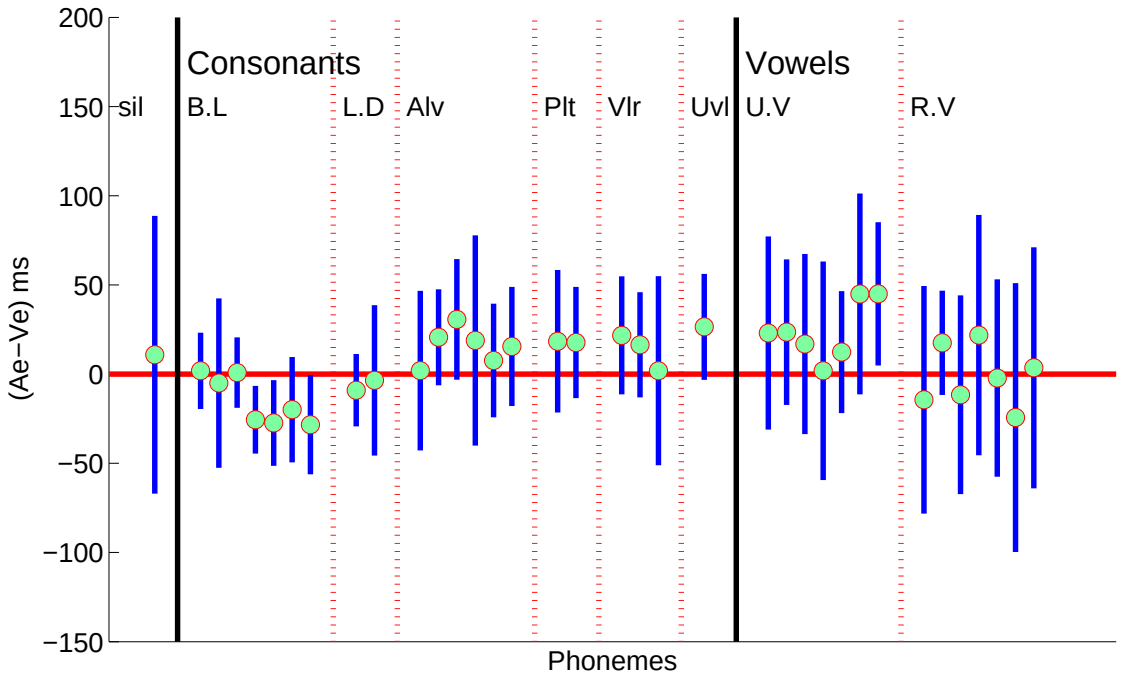

Figure 4.4: *Mean difference in ends of acoustic and visual speech segments*

4.2 Learning phoneme kinematics using EMA data

The main disadvantage of facial speech data is that, the kinematics of the invisible phonemes cannot be captured. This is because of the fact that, kinematic information about the tongue which is one of the active articulators for many phonemes, does not get captured. Alignment experiments were done to estimate the component of this missing information which can be supplemented through the addition of a tongue. The alignment experiments were performed using a data which included the tongue trajectories during phoneme articulation. This data is different from the data utilized for the segmentation experiment described in the previous section.

4.2.1 Data acquisition

The data was acquired using Electromagnetic articulography (EMA) (Hoole and Nguyen, 1999). EMA technique provides trajectory data of articulator flesh-points. It provides data comparable to that available from the well-established x-ray micro-beam system. EMA is extremely well suited to the study of coarticulation since it allows a wide range of utterances to be recorded in a single session. Sessions of 30 minutes or more are feasible. Moreover, it provides kinematic data in readily analyzable form. This should help to remedy one of the most serious failings

of instrumental studies of coarticulation, namely the small number of subjects per experiment. EMA is able to monitor the movements on the mid-sagittal plane of most of the articulatory structures that have been the focus of coarticulatory studies, i.e lips, jaw and tongue.

For this experiment, we utilized a different data with a different phonetic transcript. This data was acquired by Sébastien Demange (Demange and Ouni, 2011). It consists of trajectories of 8 flesh-points on the mid-sagittal plane and 4 flesh-points, symmetrically placed either sides of it. The flesh-point trajectories are recorded along with the acoustics while the subject was rendering speech (see Fig. 4.5). Sensors are glued to the skin at the 12 respective locations by surgical glue. Among these 12 sensors: 4 sensors are on the tongue, 4 sensors are on lips; 1 on the lower incisor (to track the jaw movement); 3 sensors, 2 symmetrically placed behind the ears, and 1 on the bridge of the nose (for the removal of any head movement). The data consists of 400 sentences which is for a total duration of about 16 minutes. The sensor trajectories are recorded at a sampling rate of 200 Hz. Wires connected to the sensors and the transmitters are present all the time during the acquisition. There might be twists and turns in the tongue which cannot be accurately calculated and eliminated from the acquired data. The overall accuracy of the acquired data gets affected by these drawbacks.

4.2.2 Feature extraction

Facial speech data and EMA data are not directly comparable. Considering this, alignment experiments were done using two sets of feature vectors extracted from EMA data alone. This way it would help in comparing the improvement of inclusion of the tongue data. The alignment experiments were done first using feature vectors having only the labial and jaw movement based features based features. Then the same experiment was done using vectors having both labial and jaw based features and tongue related features. Though tongue related features are also related to articulation, we refer to only the labial and jaw related feature as articulatory features in the following discussion. They are calculated just as in the case of facial data (see Fig. 4.5). The parameters related to the tongue are the ones which account for the movement of the tongue tip, horizontal displacement of the tongue, tongue shape, tongue height (see Fig. 4.6).

4.2.3 Results

The HMM training and alignment is done exactly in the same way as explained for the facial data. Two sets of HMMs are trained using the two feature sets extracted from EMA data. Only monophone HMMs were trained and used for segmentation. This is because of the coverage being low for a large set of triphones. The recognition criterion explained in the previous section

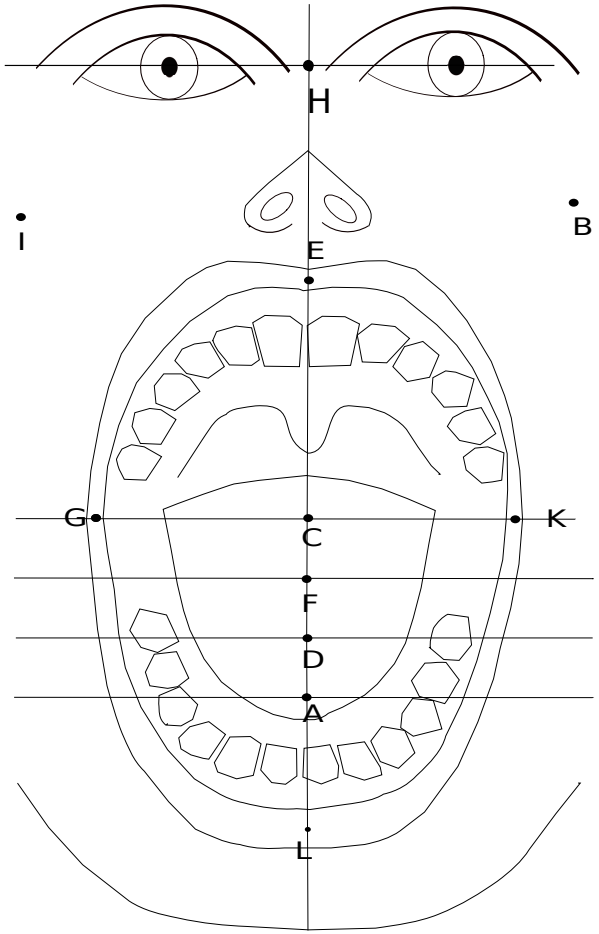

Figure 4.5: *EMA data acquisition: Location of sensors, frontal view. Lip opening and lip spread are given by the distances* $\|\vec{EL})\|$ *and* $\|\vec{GK}\|$*. Lip protrusion is given by the displacement of the center of gravity of the four points (E, G, K, L) along the normal vector to the plane formed by vectors* $\vec{EL}$ *and* $\vec{GK}$*. Figure adapted and modified from (Pfitzinger, 2005).*

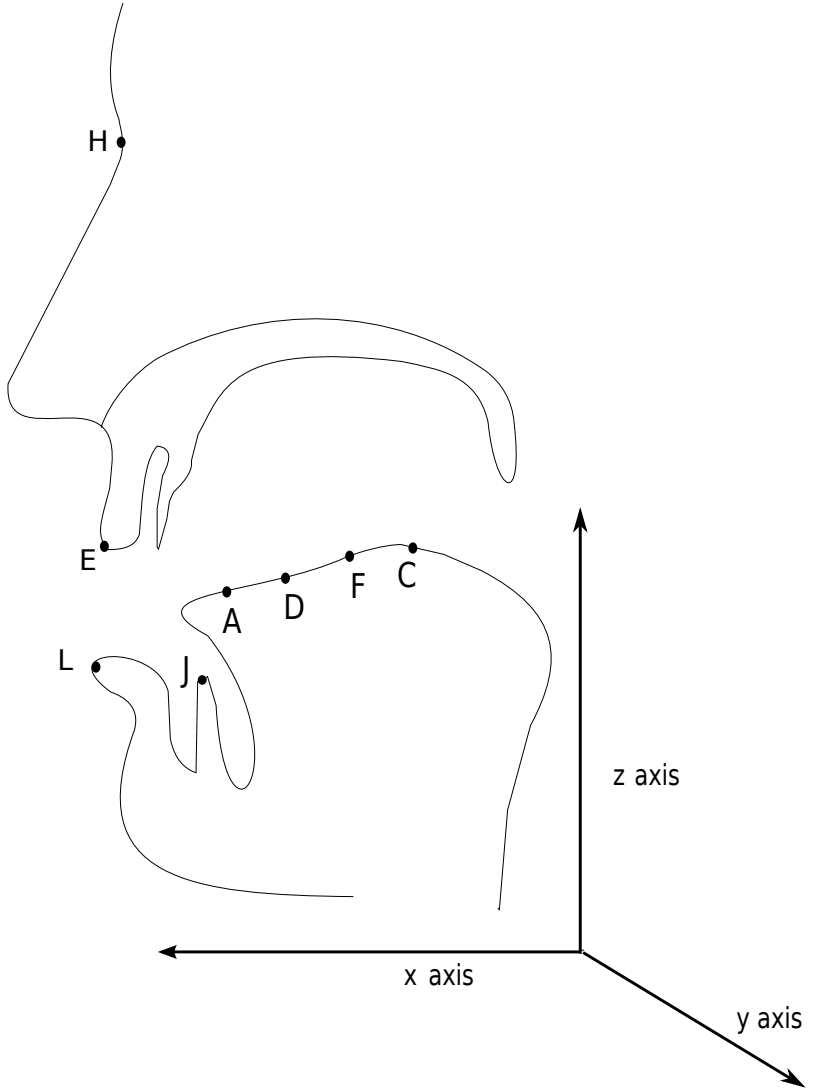

Figure 4.6: *EMA data acquisition: Sensor locations on the mid-sagittal plane. The following tongue related features are calculated: 1. Tongue tip movement,* $\|\vec{AJ}\|$*, 2. Horizontal displacement of the tongue,* $(\|\vec{JF}\|)_x$*, 3. Tongue shape,* $(\|\vec{AD}\|)_{(x,z)}$ *and* $(\|\vec{AC}\|)_{(x,z)}$*, 4. Tongue height,* $(\|\vec{JF}\|)_z$*. Figure adapted and modified from (Pfitzinger, 2005).*

is used for the analysis. The segmentation results are obtained for the two sets of HMMs. The recognition errors are determined for each phoneme class for the segmentation predicted by the two HMM sets. This is in similar lines as explained in the case of facial marker data. The results in comparison with those obtained by HMMs trained using features extracted from the facial marker data are given in figure 4.7. Facial data and EMA data have a lot of differences besides just the phonetic transcript, duration and coverage of phonemes. There are other significant differences such as the following. First, Unlike facial data where the articulation is completely uninhibited and natural, the affect of the presence of sensors on articulation cannot be completely ruled out. In addition to that, the facial deformation happening during the articulation of speech cannot be completely captured through just 5 points (4 on lips and 1 on the chin), in this respect facial data can be considered better. Besides, trajectories of just 4 points on the tongue are captured and parameters were extracted subsequently. This can not capture the complexity of the articulatory deformation of the tongue. These differences and factors account for the marginal improvement with the addition of tongue related information, which is contrary to what one would expect. Broadly, the addition of tongue features improves the alignment results for most of the phonemes which don't fall in the category of visible phonemes (see figure. 4.2). For the phonemes which fall in the category of visible phonemes, rather predictably, the addition of tongue information does not improve the recognition.

Figures 4.8 to 4.11 give the start and end statistics of the phonemes based on the alignment results without and with tongue related data to the articulatory features. Considering those phonemes for which the recognition errors have reduced with the addition of tongue data, the following observations can be made. For velars, the expectation of acoustic to visual start difference is positive, i.e. $(E(Ds) > 0)$, which indicates the co-articulation effect on their left contextual phonemes. For alveolars and dentals, the variance of the difference in acoustic and visual start (Ds) has reduced. Besides, for the phoneme /l/, the difference in the acoustic and visual ends $((E(De) < 0))$ shows an influence on the following phonemes. For other phonemes, these figures show that there is no significant change in the statistics with the inclusion of the tongue data. This can be accounted by the recognition errors, which has not improved with the addition of tongue data.

4.3 Conclusion

The results of segmentation using EMA data which includes tongue related features, in comparison of those obtained by facial features, shows only a marginal improvement. This is in agreement to the kind of result shown in (Yehia et al., 1998). We classify phonemes as visible

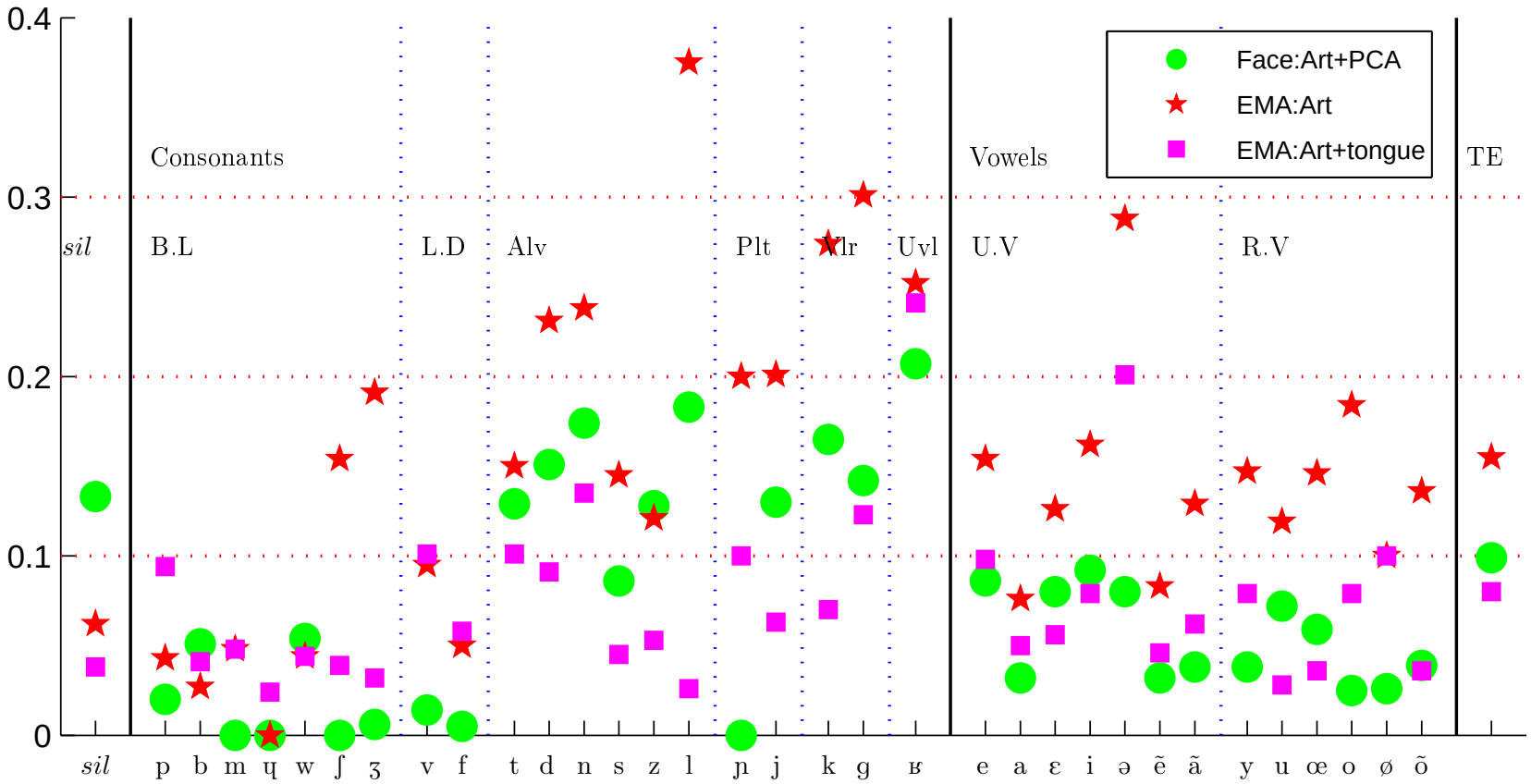

Figure 4.7: *Shows the forced alignment results using trained HMMs using different Data based features.* ***Face:PCA+Art*** *are the feature vectors extracted from the facial marker data having the four articulatory features and first 3 PCA coefficients.* ***EMA: Art*** *are the articulatory feature vector extracted from the EMA data and* ***EMA:Art+tongue*** *are the articulatory and tongue movement related feature vector from EMA data*

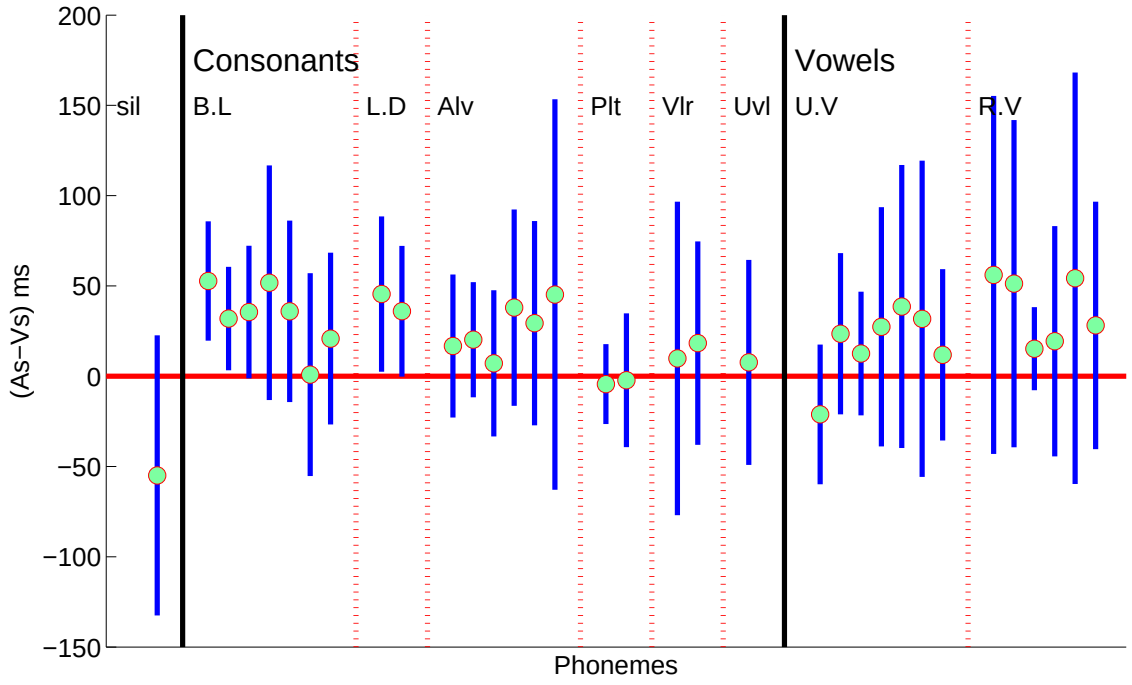

Figure 4.8: *Means and variances of the phonemes start differences calculated for the alignment based on articulatory parameters of EMA data*

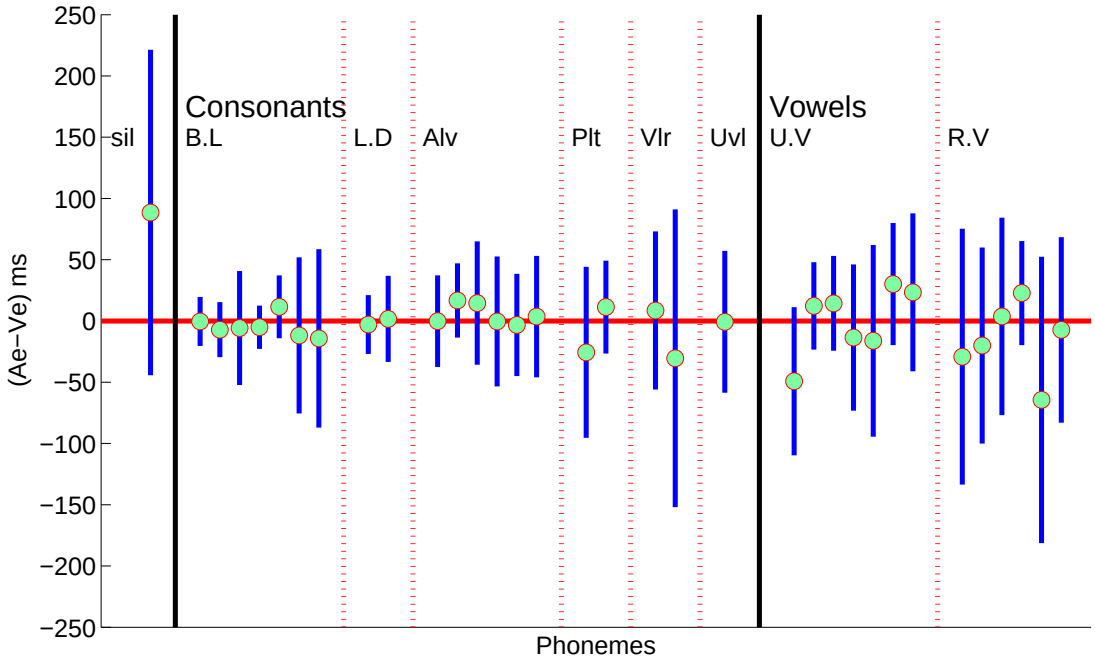

Figure 4.9: *Means and variances of the phoneme end differences calculated for the alignment based on articulatory parameters of EMA data*

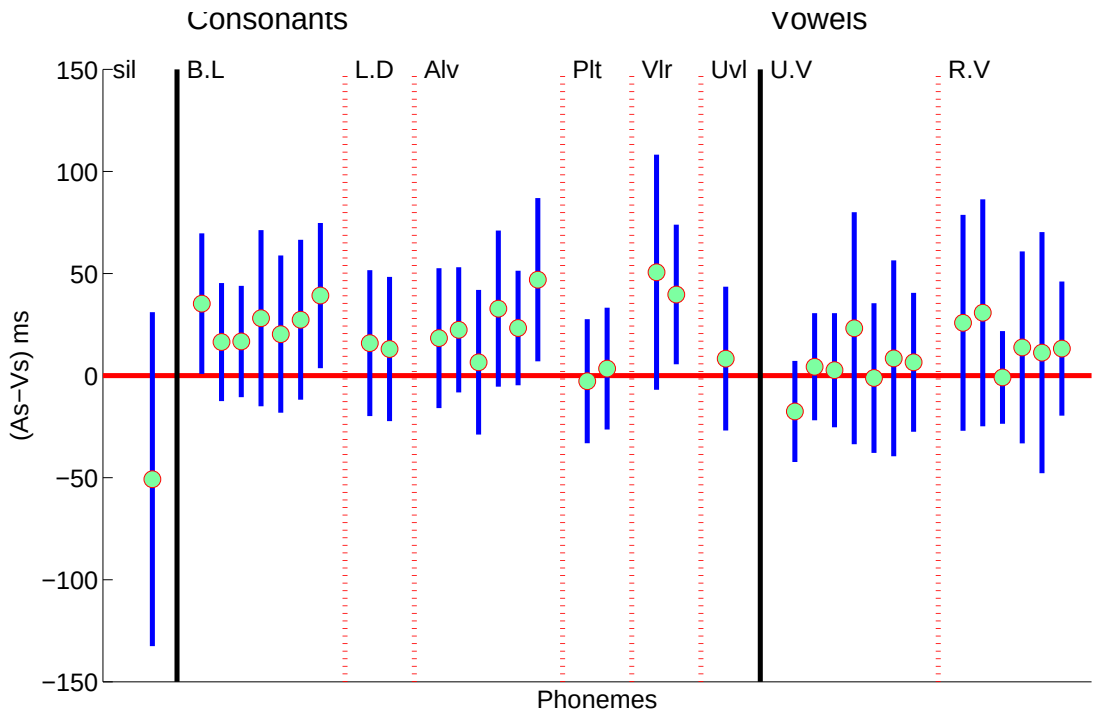

Figure 4.10: *Means and variances of the phonemes start differences calculated for the alignment based on both articulatory and tongue related parameters of EMA data*

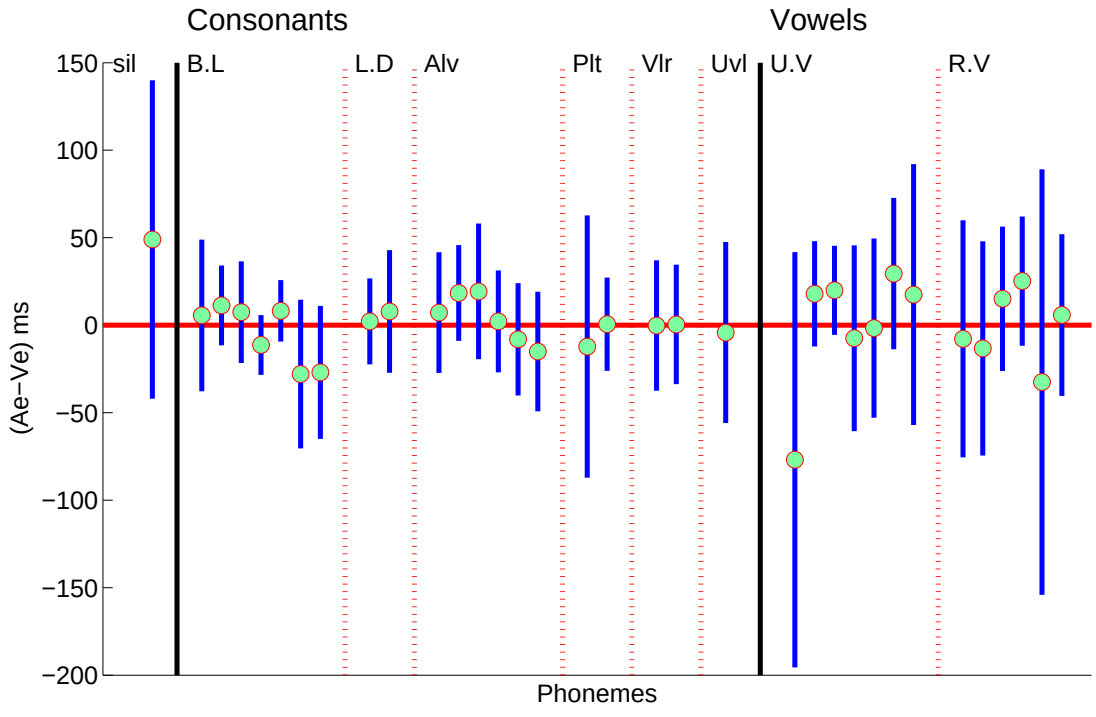

Figure 4.11: *Means and variances of the phonemes end differences calculated for the alignment based on both articulatory and tongue related parameters of EMA data*

based on these automatic segmentation results. This classification is used to analyze the perceptual evaluation results. It is useful for bringing out the correlation between objective and perceptual evaluation results, thus paving way for better objective evaluation techniques[2].

[2]Significant portion of this chapter was published as (Musti et al., 2010).

Chapter 5

Unit Selection

In the previous chapter we have presented an overview of our text to acoustic-visual speech synthesis system called ViSAC. It synthesizes speech using unit-selection and concatenation of speech segments from a pre-recorded speech corpus. Such speech synthesis systems which are based on unit selection typically have three stages. For a given text to be synthesized, the NLP module first generates the specification of the required target phoneme sequence. The specification is then converted in terms of the synthesis unit. For example, the synthesis unit in the case of our system is diphone. It is necessary that the target specification has all the important information which affects speech realization. Then, for each required target in the specification, all the candidates in the corpus are ranked based on a target cost function. This cost function is generally defined as the weighted sum of individual feature costs. At the end of this candidate ranking, for each required target in the specification, utmost a fixed maximum number of candidates are pre-selected and rest pruned. This scenario of multiple possible candidates for each required target in the sequence, defines a lattice. Finally, the sequence of those final candidates which optimizes a total cost function is selected for concatenation. This is done by the resolution of the lattice through Viterbi algorithm. The total cost function is the weighted sum of the target cost and the concatenation costs.

For all the three stages mentioned above, 'specification of targets' or 'description of candidates' is crucial. This also shows that the target feature structure and the calculation of target cost plays a central role. In the pre-selection stage, it is necessary that the ranking given to the candidates present in the corpus is consistent with the ordering based on their perceptual suitability for any required target. This is also important to ensure that no good candidates get pruned. This depends on the target cost. Besides pre-selection, target cost also influences the final selection of candidate sequence from the lattice. The set of target features and their optimum weights which define the target cost, decide the efficiency of the target cost function

and hence the synthesis performance. With respect to target cost, the following two aspects need to be explored:

- Deciding the set of target features that will be used for target specification or candidate description.
- Tuning the weights of the target features to optimize the overall synthesis performance, for a given corpus.

In addition to the target cost, the concatenation cost also needs to be considered. The concatenation cost estimates the perceptual discontinuity due to the concatenation of two candidates. The calculation of the acoustic and visual concatenation cost in our system was explained in the previous chapter. The objective of unit selection is to have a final synthesized speech which is perceptually similar to a natural speech sequence (hypothetical) rendered by the speaker. This requires at least a continuous speech without perceptible discontinuities, and constituent speech segments which are locally suitable for each required target. This requires an optimum combination of target and concatenation costs. This, indicates the need to tune the total cost function besides optimizing the total cost.

This chapter deals with these different aspects of unit selection. In the following sections, we describe experiments that were performed with the objective of optimizing the synthesis results. In the following sections, we first give an account of the set of target features in section 5.1. In section 5.2, we detail experiments that were performed to modify target feature values or design new target features for visual modality. In section 5.3, we explain a target cost tuning approach that we have developed before concluding.

5.1 Target features

At the time of synthesis, targets are specified using a set of features, generally called target features. This set of target features is generally decided based on the linguistic and phonetic studies which explain various patterns in speech. Consequently, the classically used target features include linguistic, phonetic and prosodic context. Some of these features are relevant irrespective of a language and some might be language-specific. For example, unlike phoneme voicing which is usually relevant irrespective of a language, the observation of rhythm group (RG) pattern is relevant for French. This is because in French the end of RG gives the position of the stressed syllable which is usually the last syllable of RG. Hence, the features related to RG that are relevant to French, might not be relevant or equally important for other languages. For any target or candidate, these feature values are set for both targets and candidates solely based

on the text analysis. In the case of a text to be synthesized, the description of a target in terms of these features provides an 'abstract' information about speech. The target feature cost for a particular candidate is based on the feature value of the target and that of the candidate being considered. The expectation is that same feature values account for a hypothetical similarity in the speech realization and hence also the candidate suitability.

In our system, these features describe a phoneme at various logical levels in which a sentence can be sub-divided (see Fig. 3.6). Some of the features are more specific to French language. These set of features, especially the linguistic features, are predominantly generic and can be directly applied irrespective of the corpus being used. The set of linguistic features includes phoneme number in the syllable; syllable kind; syllable position in the rhythm group (RG) and sentence; syllable number in the word, RG and sentence; word position in RG and sentence; word number in RG and sentence; RG position in sentence; proximity of the nearest left and right silence; kind of sentence.

They either have finite integral values or categorical values based on the feature. These features are either used to describe the characteristic of a target or a candidate or a contextual (left/right) phoneme or both. The phonetic features include, besides the phoneme identity, the list of features given in table 5.1. Except the phoneme identity, the other phonetic features are used to define context (left and right phoneme). This set of generic target features which are extracted through the text analysis is augmented by additional corpus-based target features. This is done to take the speaker characteristics into account which is important especially for the visual modality. Hence, the corpus specific features designed mainly account for the visual modality of speech.

5.2 Corpus based visual target features

We have described the set of generic target features in the previous section, which are generally assumed to depend solely on text analysis. The set of target features related to phonetic context also belongs to this category. The phonetic context of any particular phoneme influences its articulation significantly. This is well known as coarticulation. The degree by which a phoneme influences its surrounding phonemes or is influenced by them varies (Löfqvist, 1990). The established phonetic knowledge regarding coarticulation holds almost all the time (Ladefoged, 1982; Ladefoged and Maddieson, 1995). Hence, these target features and their values for different phonemes are usually based on the characterization defined by phoneticians that is found in the literature. Hence, their values are set based on the information extracted through text analysis. However, the phonetic context also varies significantly based on the speakers' articulatory prefer-

Table 5.1: These set of features define the phonetic context of a phoneme, target or candidate. These feature values either describe previous or following phoneme. The target feature costs for these features are binary valued functions taking either 0 or 1 based on whether the feature values being compared are same or different respectively.

Feature Name	Possible values
Voicing	voiced, unvoiced
Kind	vowel, consonant, semivowel
Place of Articulation	bilabial, labiodental, inter-dental, alveodental, alveolar, post-alveolar, palatal, post-palatal, prevelar, velar, post-velar, uvular, laryngeal, lateral
Manner of Articulation	Oral, nasal, plosive, fricative, liquid, semi-plosive
Lip Shape during articulation	spread, protruded

ences and idiosyncrasies. Due to the usage of a recorded audio-visual corpus, in case the speaker has any peculiar articulation, it might be visually or acoustically perceived in the synthesized speech and present some incoherence. For example, let us assume that candidates are being looked up for a target phoneme whose left contextual phoneme is considered to have lip protrusion during its articulation. Then obviously, those candidates whose left contextual phoneme is considered to have a lip protrusion during its articulation will get higher ranking. If this target contextual phoneme is actually articulated differently and not actually protruded, then selecting a candidate with a protrusion left contextual phoneme might be inappropriate. This kind of categorization might slightly vary from person to person and it is well known (Johnson et al., 1993; Raphael and Bell-Berti, 1975; Maeda, 1989). Hence, in case these feature values have any inconsistency in comparison with the actual characteristic in the corpus, it will be visible in the synthesized speech. We have performed two experiments which aim at a phonetic context adaptation that is based on the characteristics observed in the corpus. They can be divided into the following two categories:

- Changing target feature values for some phonemes based on the articulatory characteristics estimated from the corpus. We refer to this approach as phonetic category modification.

- Replacing categorical phonetic target features, by real valued target features to represent corpus specific characteristics. These features encode the same information accounted by the categorical features, with higher precision. We refer to this approach as continuous visual target cost.

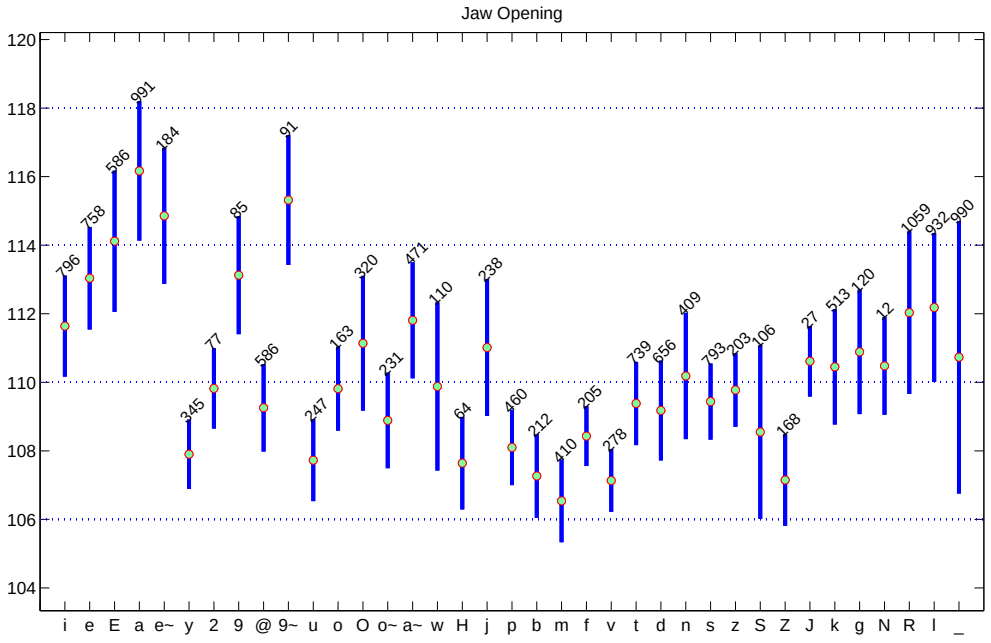

Figure 5.1: *Jaw Opening statistics. Each segment represents a phoneme, centered at the mean and its length being twice the standard deviation. The number of occurrence of each phoneme is presented.*

In the following subsections, we describe these experiments. The modified feature values or introduced features are those which mainly characterize the visual modality of speech. Hence, we refer to them as visual target cost. The main goal is to see whether these experiments improve the performance of selection and consequently of synthesis. The objective evaluation results of these two methods are then presented in subsequent subsections.

5.2.1 Phonetic category modification

All the target features which provide the information related to phonetic context are categorical (see Table 5.1). The corresponding phonetic feature costs are binary; which take 0, when the target and candidate feature values are same and 1, when they are different. Among these target features, two features account for the patterns in visual speech animation. They are 'Place of articulation' and 'Lip shape during articulation'. We would refer to the latter feature as 'Lip Shape'. 'Place of articulation' information is encoded only for labial phonemes and also their place of articulation is visibly unambiguous. Hence, we focus on 'Lip Shape'.

We want to determine the characteristic lip shapes of phonemes as observed and directly measurable from the recorded audiovisual corpus. In case the observed 'Lip Shape' is different from the expected classical categorization, the category is modified accordingly. This information will be used to redefine this feature's values while specifying targets and describing candidates

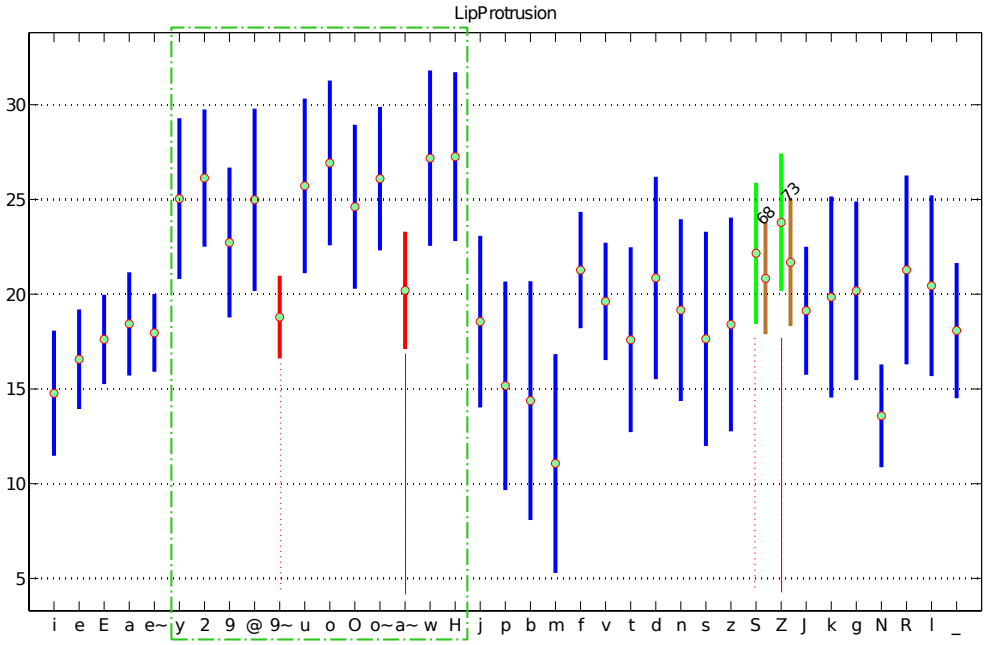

Figure 5.2: *Lip Protrusion statistics. The phonemes of interest are framed: the 'protruded' phonemes are {y, ø, œ, ə, œ̃, u, o, õ, ɔ , ã, w, ɥ }. The segments plotted in red, green and brown seem to violate the general pattern recalculated with candidates without a 'protruded' context. The segments plotted in red correspond to the phonemes whose category was modified. The brown and green segments are of those phonemes where statistics were recalculated with candidates without 'protruded' context.*

more accurately. The expectation was that their synthesized visual speech component would be more similar to the real visual speech after the changes. This modification of the phonetic context should modify the visual target cost, which is a part of the target cost (TC). The visual target cost of a phoneme (left or right phoneme of a diphone) is calculated by summing the visual feature differences of the left and the right contextual phonemes.

We performed a statistical analysis of the articulatory features. These set of articulatory features included lip protrusion, lip opening, lip spreading and jaw opening (see Fig. 3.4) (Robert et al., 2005). The statistics were calculated by considering the articulatory feature vectors at the center of the phoneme articulation. This is also the place of concatenation in the visual and acoustic domain. The statistics of the phonetic articulatory features are shown in figure 5.1 to 5.4. We considered the mean, variance and the number of occurrence of each phoneme. For any given phoneme, the lip shape can be either 'Protruded' or 'Spread', or might not have any typical shape in which case we classify as 'not protruded and not spread' which we refer to as simply 'none'. The range of articulatory feature statistics for each of these categories is determined first. This is depends on the pattern that majority of phonemes belonging to each category seem to follow. Each phoneme category is re-examined based on these intervals thus

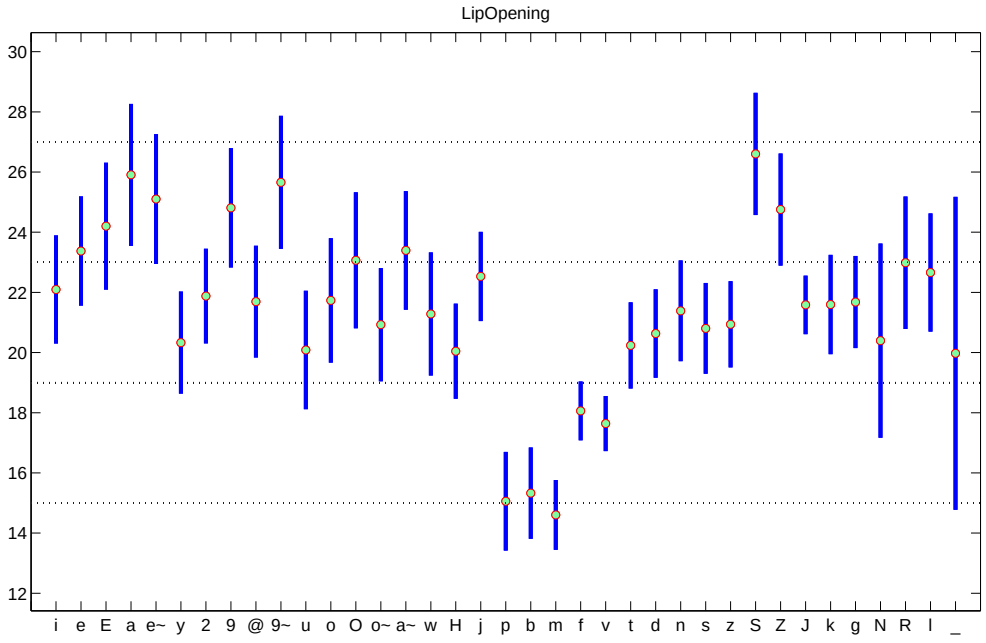

Figure 5.3: *Lip Opening statistics.*

determined. We looked more closely at LipProtrusion and LipSpread as others are related.

Typically by classical phonetic knowledge, the set of phonemes which included {y, ø, œ, ə, œ̃, u, o, õ, ɔ , ã, w, ɥ } was classified as 'protruded' and the set of phonemes which included { i, e, a, ɛ, ẽ} was categorized as 'spread' phonemes . All the other phonemes were considered as 'not spread and not protruded' based on the shape of the lips. This categorization generally holds. Nevertheless, we can observe that some phonemes need to be reconsidered. For this purpose and to be more accurate, the coarticulation affects of the surrounding phonemes should be removed. In fact, if one of the neighboring phonemes is protruded, for instance, it is very likely that the surrounded phoneme will be protruded too, even if it is not its main articulatory characteristic, because of coarticulation. Therefore, for phonemes whose visual articulation seemed to be different from their initial classification, their articulatory feature statistics were recalculated by considering a subset of phoneme instances in the corpus. For example, the phoneme /f/ seemed to be 'spread' unlike its classical phonetic classification of 'not spread'. Thus, only its occurrences in the corpus without spread phonemes in its neighborhood were taken into account. Its articulatory feature statistics were recalculated to confirm its effective visual articulation. The following set of phonemes were considered for recalculation to check if their effective articulation is 'spread': {f, v, t, d, n, s, z, ɲ, k, g, ŋ }. For the two phonemes {ʃ and ʒ}, the articulatory feature statistics without rounding context was recalculated. These statistics were recalculate to ensure that the observations are not due to the contextual influence

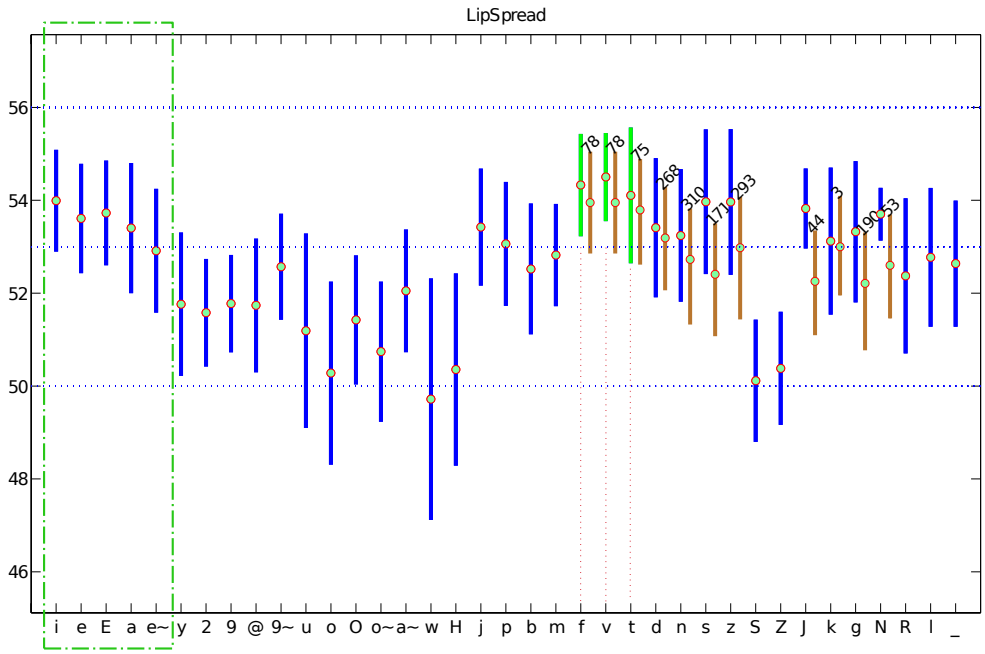

Figure 5.4: *Lip Spread statistics. The phonemes of interest are framed: the 'spread' phonemes are { i, e, a, ɛ, ɛ̃}. The brown and green segments seem to violate the general pattern and so their statistics were recalculated with candidates without 'spread' context. The segments plotted in green correspond to the phonemes whose category was modified.*

but representative of the phoneme articulation itself. Initially, the sets of phonemes {f, v, t}, {ʃ, ʒ} and {ã, œ̃} were considered as 'none', 'none' and 'protruded' respectively. However, based on the statistics and the observation of the data, we found out that the strategy of our speaker is quite different from this definition. For this reason, we modified the articulatory target features for these sets phonemes to 'spread', 'protruded' and 'none' respectively.

In subsection 5.2.3, we present an evaluation where we compared the synthesis using the initial articulatory description (IPD) and the modified phonetic description (MPD).

5.2.2 Continuous visual target cost function

In the previous subsection, we explained the re-classification of phonetic characteristics into distinct categories from the statistics of the articulatory features. The goal was to adapt the classification to the real ones based on the corpus used. But one can observe that it is not easy to take a discrete distinct decision from these statistical values. So the visual target cost component has to be formulated as a real value in the range $[0, 1]$ rather than binary value. The articulatory characteristics should be considered as continuous. So the visual target cost component has to be formulated as a real value in the range $[0, 1]$ unlike binary value. For calculating the continuous target cost we used the articulatory feature statistics calculated as

explained in the previous subsection. We explored two different formulations of continuous visual target cost. First formulation is based on a work done by Mattheyses et al. (2010) which uses contextual phoneme difference. The second formulation is based on an approach that we developed, which is based on contextual significance. The articulatory feature statistics are represented by μ_{ij} and σ_{ij} to represent the mean and variance of the phoneme (index i) and using the articulatory feature (index j).

5.2.2.1 Visual target cost function based on contextual phoneme difference

In (Mattheyses et al., 2010), the authors used shape and texture parameters extracted by applying Active Appearance Models on 2D facial images of speech animation. We tried to apply the same logic for the calculation of the continuous target cost using articulatory features. In this formulation, the calculation of visual target cost is done as follows: Two phonemes are considered similar in terms of their visual representation, if their mean representations are alike and, in addition, if these mean representations are sufficiently reliable (i.e. with small summed variations). Two matrices were calculated, which express for each phoneme pair (p, q); the difference between their mean representations D^{μ}_{pq} and the sum of the variances of their visual representation D^{σ}_{pq}, respectively:

$$D^{\mu}_{pq} = \sqrt{\sum_{j} (\mu_{pj} - \mu_{qj})^2}$$

$$D^{\sigma}_{pq} = \sum_{j} \sigma_{pj} + \sum_{j} \sigma_{qj}$$

Scaling both matrices between zero and one gave $D^{\mu'}_{pq}$ and $D^{\sigma'}_{pq}$, after which the final difference matrix was calculated:

$$D_{pq} = 2D^{\mu'}_{pq} + D^{\sigma'}_{pq}$$

Matrix D_{pq} is used to calculate the visual target cost during selection.

5.2.2.2 Visual target cost function based on contextual significance

In the previous method, the point of emphasis was centered on the differences in contextual phonemes. It doesn't take into account the nature of the main target phoneme. For each phoneme, the feature with least variance is the one which gets least modified due to coarticulation and the features with higher variance get affected more due to coarticulation. Thus, obtaining similar context is important for features which get more influenced due to coarticulation. We applied this principle for the calculation of contextual phoneme difference $D_{pq}(i)$ as a function

of the central target phoneme i which is being looked up in the corpus. The following notation is assumed: p is the contextual phoneme (left or right) of phoneme i in the target utterance and q is the contextual phoneme of the candidate for i. The difference of the mean of the contextual phoneme was weighted by the variance of the target phoneme:

$$D_{pq}(i) = \sum_j w_{ij} |\mu_{pj} - \mu_{qj}| \tag{5.1}$$

$$w_{ij} = \frac{\sigma_{ij}}{\sum_j \sigma_{ij}}$$

$D_{pq}(i)$ is scaled between zero and one. This gives the distance between contextual phonemes as a function of the phoneme i for which, the proximate context is being looked up during the selection process. The weight w_{ij} gives the relative importance of the component j with respect to the other components. Higher the variance σ_{ij}, higher the weight on the contextual difference for the component j. Thus, w_{ij} reflects the fact that context has important impact on these components with higher variance.

5.2.3 Objective evaluation of synthesis results

In this subsection we describe the objective evaluation done to compare the various visual target costs. For the purpose of evaluating the synthesis results, we used a method based on leave-one-out cross-validation technique. We synthesized each of the sentences in the corpus, a total of 319 sentences. This is done by excluding the sentence being synthesized from the selection corpus. Each of the synthesized sentences are compared with the real sentences. The advantage of this method is that it avoids building a specific test corpus for evaluation. However, we marginally reduce the choice of selection, by excluding some diphones from the selection process.

After synthesizing a given sentence, all the half-phones (two half-phones in each diphone) of the synthesized sentence and the actual sentence were re-sampled individually to make the number of visual samples equal in both the real and synthesized sentences (see Fig. 5.5). This was done using a simple linear interpolation of the 12 PCA coefficients. After this, the Pearson's correlation coefficients between 12 PCA coefficients of all the synthesized sentences and the real sentences actually present in the corpus was determined. Similarly, Pearson's correlation coefficients between 4 articulatory parameters was also determined. The root mean square error (RMSE) between articulatory feature and PCA coefficient trajectories of the synthesized and the real sentences present in the corpus was determined.

If x_d and y_d are the sequences of the d^{th} PCA coefficient of a real and synthesized sentence

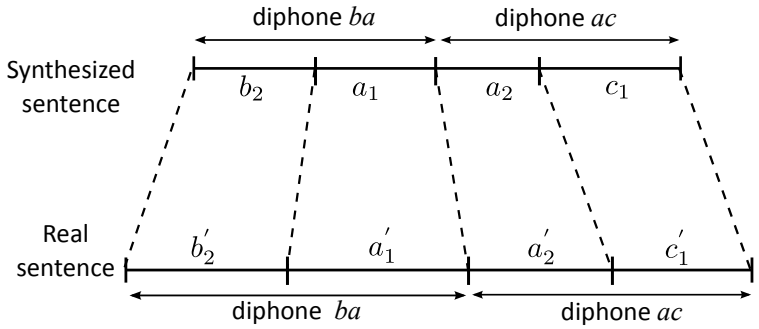

Figure 5.5: *Adjusting diphone lengths. Each of the corresponding half-phones which are part of the diphones in the synthesized and real sentences are re-sampled through linear interpolation to make the number of visual samples equal.*

having n samples:

- The Pearson's correlation coefficient is calculated as follows:

$$r_{x_d y_d} = \frac{n \sum_i x_d(i) y_d(i) - \sum_i x_d(i) \sum_i y_d(i)}{\sqrt{n \sum_i x_d(i)^2 - (\sum_i x_d(i))^2} \sqrt{n \sum_i y_d(i)^2 - (\sum_i y_d(i))^2}} \tag{5.2}$$

- The Root Mean Squared Error (RMSE) is calculated as follows:

$$rmse_{x_d, y_d} = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (x_d(j) - y_d(j))^2} \tag{5.3}$$

Though it is almost impossible to have a perfect correlation between the real and synthesized sentence, it seems to be a reasonable assumption that the trajectories for two diphones selected with similar phonetic context and linguistic description would be significantly correlated. For the visual target cost, we performed objective evaluation of the visual speech animation alone. This was based on the assumption that the visual speech animation would be strongly correlated with the underlying acoustic speech. Besides, the features modified account predominantly for the visual modality of speech unlike some others like phoneme articulation, voicing which account for the acoustics of speech. An example of the trajectories of the first principal component of a synthesized sentence and the corresponding real sentence are shown in figure 5.6.

Evaluation results

Based on the above explained objective evaluation technique the performance of the various visual target cost techniques were determined (See tables 5.2 and 5.3). The target cost techniques with the binary visual target cost components (see section 5.2.1): Initial articulatory

description (IPD) and Modified phonetic description (MPD) performed comparable to each other ($r_{x_d y_d} = 0.813$ for PC 1). Similarly, the two continuous visual target costs; contextual phoneme difference based approach (CPD) and phoneme difference based on contextual significance (PDCS) performed comparable to each other ($r_{x_d y_d} = 0.816$ for PC 1). The continuous visual target costs gave marginally better results consistently compared to the binary visual target cost approaches even when different weights for the visual target cost component were used. This is also apparent when observing the performance with respect to articulatory features. In fact, the correlation for the first two methods IPD and MPD is 0.70 and it increases up to 0.72 for the CPD and PDCS for jaw opening (see table 5.2). Table 5.3 shows the RMSE between real and synthetic trajectories for the articulatory features. The RMSE is almost the same for the 4 methods. We should notice that each of the examined methods affects the ranking of the selected candidates though it is not that obvious that there are differences between them. We should emphasize that the relative importance of this examined visual target cost component in the overall target cost is 1%, as we have a large set of features. Therefore this can explain this marginal variation in the performance.

Hence, these results indicate that a continuous target cost component represents the differences between phonemes better, optimizing the synthesis performance for particular corpus than discrete binary target cost components has to be contemplated. Given the limited generalizing power, for a corpus of small size and without a very well balanced diphone coverage in the corpus, the categorical target cost based on classical knowledge can be considered sufficient. One should observe that the objective evaluation used in this work is purely visual.

Examining the results of the objective evaluation presented here, it can be said that they are quite good. The overall correlation is quite high. In addition, the RMSE is very low and acceptable. In fact, the jaw opening RMSE is around $2mm$, lip opening ($2.7mm$), lip spreading ($1.38mm$) and lip protrusion is $4mm$. This is a good indication that our synthesis method provides similar trajectories to those of real sentences. This is quite interesting, as we know that the purpose of synthesis is not to generate the exact speaker articulation (unlike acoustic-to-articulatory inversion). As natural speech realization is variable and so good synthesis can also be obtained by different trajectories which don't exactly match with one real reference. But as our system takes into account the specificity of the speaker into account, we manage to obtain a similar result which is closer to the speaker's articulation. Thus, it seems that our acoustic-visual synthesis, based on the main idea of considering the speech signal as bimodal, was able to capture the speaker specific articulation finely. This can be clearly seen in Figure 5.6. It clearly indicates that it might improve the synthesis results if the target features are

modified/optimized to take any particular corpus they describe.

PC	IPD	MPD	CPD	PDCS
1	0.813	0.813	0.816	0.816
2	0.715	0.715	0.719	0.720
3	0.726	0.725	0.729	0.729
JO	0.708	0.708	0.728	0.728
LP	0.694	0.693	0.698	0.698
LO	0.671	0.670	0.689	0.689
LS	0.636	0.636	0.640	0.640

Table 5.2: Correlation coefficients between the real and synthesized trajectories of first 3 principal component coefficients and the three articulatory features by various target cost strategies. IPD: initial phoneme description, MPD: Modified phoneme description, CPD: contextual phoneme difference, PDCS: phoneme difference based on contextual significance. The articulatory features: JO (jaw opening), LP (lip protrusion), LO (lip opening) and LS (lip spreading). The first four principal components account for about 58%, 24% and 7% respectively.

PC	IPD	MPD	CPD	PDCS
1	7.86	7.86	7.78	7.77
2	6.67	6.67	6.63	6.62
3	5.67	5.67	5.64	5.64
JO	2.11	2.11	2.06	2.06
LP	4.04	4.04	4.02	4.02
LO	2.70	2.70	2.63	2.63
LS	1.38	1.38	1.37	1.37

Table 5.3: Root Mean Square Error (RMSE) in millimeters between the real and synthesized trajectories of the four articulatory features (same notations as table 5.2).

5.3 Target feature selection and weight tuning

The key to the synthesis of 'natural' sounding speech is the assignment of a target cost which is correlated to human perception. This is important not only for the pre-selection of appropriate candidates from a large corpus but also for the selection of the final candidate sequence for synthesis. The set of target features and their optimum weights affect the performance of the target function. Once the set of target features is decided, the target feature weights are tuned such that the overall synthesis performance is the best possible with the corpus being used.

We developed an iterative algorithm to simultaneously perform redundant feature elimination and weight tuning. The algorithm which is applicable to unit selection based speech synthesis in general, is presented in the context of Audio-Visual speech synthesis. A target cost function is evaluated based on the comparison of its candidate ranking and the ordering given by an objective dissimilarity measure comparing two speech segments. This target cost evaluation is similar to the Minimum Selection Error approach presented in (Latacz et al., 2011). It is

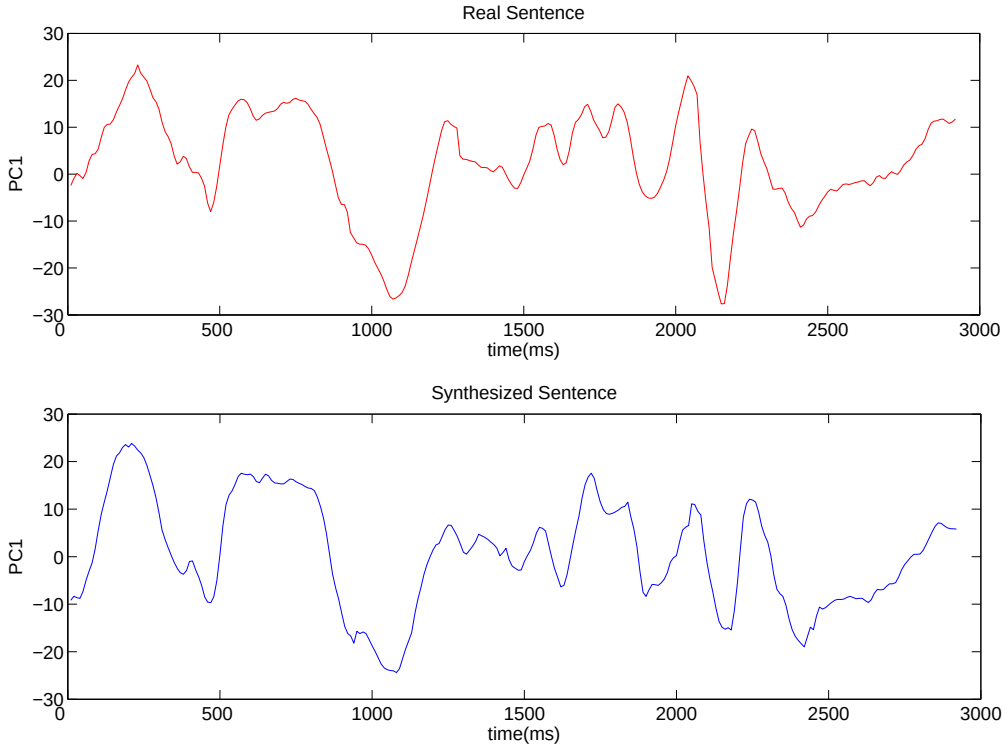

Figure 5.6: *Resampled Synthetic and Real trajectories for the first principal component for the sentence "Sur ces mots, elle sortit vivement de la pièce." with the following phoneme sequence "sil s y ʁ s e m o sil ɛ l s ɔ ʁ t i v i v ə m ã d ə l a p j ɛ s sil" (see Fig. 5.5). The Pearson correlation for the first principal component was* 0.89.

generally possible that during weight tuning some target features are assigned negligible weights. This is implicitly feature elimination. Unlike this implicit feature elimination, we perform explicit feature elimination and update weights of all the other retained features, both simultaneously in each iteration.

5.3.1 Unit selection and concatenation

We briefly revisit the unit selection framework for speech synthesis. A typical TTS (text to speech synthesis) algorithm can be broadly divided into two steps, generation of specification and the actual synthesis. This division is made to separate the steps which perform a target cost calculation from those which do not. In the first stage, the text to be synthesized is analyzed. This stage produces the specification of the phoneme sequence to be synthesized $t_1^n = (t_1, ...t_j, ...t_n)$, n phonemes starting from 1, for the input text. The second stage does the actual synthesis of the required phoneme sequence in two steps, pre-selection and final selection through lattice resolution. This second synthesis stage depends on the target cost calculation for its synthesis performance. The target cost calculation is done by the comparison of target specification to the candidate description in the corpus. The set of candidates which

are 'perceptually' similar are pre-selected for the final search based on this target cost. A general target function is calculated as follows:

$$C(t_i, u_i) = \sum_{\rho=1}^{F} w_\rho C_\rho(t_i, u_i) \tag{5.4}$$

where, t_i, u_i are the target and a candidate; F is the number of target features; $C_\rho(t_i, u_i)(\rho = 1, ..., F)$ is the target feature costs between the elements of the target and candidate feature vectors; w_ρ is the weight of a feature ρ:

The selection among the set of pre-selected candidates is operated by resolution of a lattice of candidates using the Viterbi algorithm. The result of this selection is a path in the lattice of candidates which minimizes a weighted linear combination of three costs: the target cost ($\sum_{i=1}^{n} C(t_i, u_i)$), the acoustic join cost ($\sum_{i=2}^{n} C^{aj}(u_{i-1}, u_i)$), and the visual join cost ($\sum_{i=2}^{n} C^{vj}(u_{i-1}, u_i)$), that is

$$C^T(t_1^n, u_1^n) = \min_{u_1,\ldots,u_n} \begin{cases} w \sum_{i=1}^{n} C(t_i, u_i) & + \\ w_{aj} \sum_{i=2}^{n} C^{aj}(u_{i-1}, u_i) & + \\ w_{vj} \sum_{i=2}^{n} C^{vj}(u_{i-1}, u_i) & \end{cases} \tag{5.5}$$

where w, w_{aj} and w_{vj} are weights for the component target cost, acoustic join cost and visual join cost. We choose these weights as explained in (Toutios et al., 2011) (see section 6.1.2).

An ideal target cost function

The usage of target cost function is to rank candidates in the order of their suitability to fit a target position during synthesis. Each candidate is assigned a cost (positive real number) by the target cost function, lower the cost better suitable is the candidate for a target position. If we assume that there is a metric to measure the perceptual dissimilarity between a target and a candidate, then ideally, the ranking of candidates based on their target costs should be the same as that of the ordering based on their perceptual dissimilarities to the target.

At the time of synthesis, the target specification only has the target feature description, but no acoustic or visual speech realization. So, the decision is made based on the target cost. Hence, an optimum target cost function is very important for good synthesis results. A good set of target features and well tuned weights define a good target cost function. The following section presents a simple and robust iterative algorithm to simultaneously eliminate redundant features and tune the weights of other target features.

5.3.2 Target feature selection and weight tuning

The algorithm to be described alleviates the problem of redundancy and noise that is set in due to the exhaustive set of features considered. Its importance is also due to the fact that, with a large set of features, it is practically infeasible to have a corpus which covers all the feature combinations possible. The algorithm uses the corpus, for which we have both actual speech realizations and target feature descriptions for each of the candidates present in it.

Since for any speech segments, there are possible variants which are perceptually considered good alternatives. But, it is practically impossible to rank candidates in terms of their absolute perceptual quality with respect to any target. Being 'similar' to an already existing speech unit is a reasonable way to say how well will a candidate fit in a 'target' position. If we device a way to measure the dissimilarity between two units, it can be used on the candidates in the corpus. They have both the target feature description and speech realization available. The comparison between the ordering obtained by this measure versus the ranking using the target cost can be used to evaluate the target function. In the following paragraphs, we define two things necessary for the evaluation of a target function: disorder with respect to a target cost function and dissimilarity between two speech realizations.

5.3.2.1 Disorder

The disagreement in the ranking of candidates given by the target cost function versus the ordering by dissimilarity measure, needs to be quantified. With respect to a particular target t whose speech is available, the candidate ranking based on the target cost function should be in agreement with their dissimilarity based ordering. We refer the ordering based on the target cost as ranking. Consider a target t and two candidates u and v. With respect to the target t, let their dissimilarity measures be $D(t,u)$ and $D(t,v)$, and their target costs be $C(t,u)$ and $C(t,v)$. Then for an ideal target cost function, one of the following three conditions should be true:

1. $C(t,u) < C(t,v) \quad \Leftrightarrow \quad D(t,u) < D(t,v)$
2. $C(t,u) < C(t,v) \quad \Leftrightarrow \quad D(t,u) < D(t,v)$
3. $C(t,u) < C(t,v) \quad \Leftrightarrow \quad D(t,u) < D(t,v)$

The dissimilarity measure is based on the comparison of two speech realizations. We assume that similar speech realizations are perceptually similar. This assumption implies that the dissimilarity gives an accurate estimate of the perceptual suitability of a candidate. So through

Target Cost based ranking	Dissimilarity based ordering		Disorder calculation
	Ideal scenario	Real scenario	
$C(t,c_1)$	$D(t,c_1)$	$D(t,c_1)$	c_1 : $\delta_t(c_1,c_2)+\delta_t(c_1,c_3)+\delta_t(c_1,c_4) = 0+0+0$
$C(t,c_2)$	$D(t,c_2)$	$D(t,c_4)$	c_2 : $\delta_t(c_2,c_1)+\delta_t(c_2,c_3)+\delta_t(c_2,c_4) = 0+\lvert D(t,c_2)-D(t,c_3)\rvert+\lvert D(t,c_2)-D(t,c_4)\rvert$
$C(t,c_3)$	$D(t,c_3)$	$D(t,c_3)$	c_3 : $\delta_t(c_3,c_1)+\delta_t(c_3,c_2)+\delta_t(c_3,c_4) = 0+\lvert D(t,c_3)-D(t,c_2)\rvert+\lvert D(t,c_3)-D(t,c_4)\rvert$
$C(t,c_4)$	$D(t,c_4)$	$D(t,c_2)$	c_3 : $\delta_t(c_4,c_1)+\delta_t(c_4,c_2)+\delta_t(c_4,c_3) = 0+\lvert D(t,c_4)-D(t,c_2)\rvert+\lvert D(t,c_4)-D(t,c_3)\rvert$

Table 5.4: This table illustrates the idea of comparison of a dissimilarity measure based ordering and the ranking assigned based on the target cost. A target t and four candidates $\{c_1,c_2,c_3,c_4\}$ are assumed. It is assumed that for the target and the candidates, the speech realization is available for comparison. $D(t,c_i)$ is the dissimilarity between the speech realizations of the target t and candidate c_i, which is a symmetric function. $C(t,c_i)$ is the target cost between the target specification of t and candidate c_i. For the given target and with respect to each available candidate, the dissimilarity based ordering of candidates and the target cost based ranking is compared to calculate the disorder. The total disorder is the sum of the fourth column.

the dissimilarity measure we are expressing the difference in their speech realizations. Our approach is based on this idea that the ordering given by an ideal target cost function should agree with the ordering given by this dissimilarity measure. During pre-selection, the target cost function assigns a ranking to the available candidates, for pruning the less suitable candidates. For this reason, we refer to the target cost based ordering as ranking. Unlike some systems we don't train the target cost function to computes the dissimilarity (Hunt and Black, 1996). We only focus on the candidate ordering given by the target cost function. The above three conditions state that, the comparison of two candidates for a target position based on their target costs would be similar to that based on their dissimilarity to the target, if the target was to have a speech realization available (hypothetical). We denote the above three conditions by the following:

$$C(t,u) * C(t,v) \quad \Leftrightarrow \quad D(t,u) * D(t,v) \tag{5.6}$$

Where, $* \in \{<,=,>\}$.

We define the disorder with respect to this target and the two candidates as follows:

$$\delta_t(u,v) = \begin{cases} 0 & \text{if condition (5.6) holds} \\ |D(t,u)-D(t,v)| & \text{else} \end{cases} \tag{5.7}$$

The above mentioned explanation is illustrated in table 5.4.

For each of the phonemes p in the phoneme set, let U_p be the complete set of candidates in the corpus with that phonemic label. Using leave-one-out technique, considering each of the elements from this set as a target and all the others as candidates, the *total disorder* for that phoneme is calculated for a particular target cost function as follows:

$$\Delta = \sum_{t} \sum_{(u,v)} \delta_t(u,v) \tag{5.8}$$

Where, $u, v, t \in U_p$ and $t \neq u \neq v$. In the following sections we refer to this *total disorder* as simply *disorder*.

5.3.2.2 Dissimilarity of two units

We take a dissimilarity measure similar to that in (Latacz et al., 2011) for the acoustic modality. Here, we describe a function that we have used to compare two speech segments. It gives an estimate of their dissimilarity. We considered four components to constitute the dissimilarity measure $D(u,v)$ between units u and v of a particular phoneme p as follows:

$$\begin{aligned} D(u,v) = \quad & w_{dur} D^{dur}(u,v) + w_{ac} D^{ac}(u,v) + \\ & w_{vs} D^{vs}(u,v) + w_{f0} D^{f0}(u,v) \end{aligned} \tag{5.9}$$

D^{dur}, D^{ac}, D^{vs} and D^{f0} are the components in terms of the duration, acoustic speech, visual speech and f0 of the units and w_{dur}, w_{ac}, w_{vs} and w_{f0} are the weights given to these respective components. The duration dissimilarity D^{dur} is calculated as the difference between the durations of the two units v and u, dur_u and dur_v respectively and normalized to make the value lie in the range [0,1]. $dur_{min}(p) = \min_{u,v \in U_p} |dur_u - dur_v|$ and $dur_{max}(p) = \max_{u,v \in U_p} |dur_u - dur_v|$, which are the maximum and minimum duration differences among the units of phoneme p. Then, the duration dissimilarity component is calculated as follows:

$$D^{dur}(u,v) = \frac{|(dur_u - dur_v)| - dur_{min}(p)}{dur_{max}(p) - dur_{min}(p)} \tag{5.10}$$

For the other three components; acoustic, visual and f0; the RMSE (root mean squared error) is calculated between two trajectories of respective features by making the duration or number of samples N equal by simple linear interpolation.

$$d^{rmse}(u,v) = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (x_u(j) - x_v(j))^2} \tag{5.11}$$

The features used for visual and acoustic dissimilarity measure are PCA coefficients and

MFCC as explained in section 5.3.3.

$d_{min}(p) = \min_{u,v \in U_p} d^{rmse}(u,v)$ and $d_{max}(p) = \max_{u,v \in U_p} d^{rmse}(u,v)$, which are the maximum and minimum RMSEs among all the units of phoneme p. The RMSE is normalized similar to D^{dur} to make the value lie in the range $[0,1]$ using $d_{min}(p)$ and $d_{max}(p)$:

$$D^{rmse}(u,v) = \frac{d^{rmse}(u,v) - d_{min}(p)}{d_{max}(p) - d_{min}(p)} \tag{5.12}$$

5.3.2.3 Primitives of the algorithm

The main idea behind the algorithm to be described is that, each target feature has some contributing information which gets reflected in speech. If a useful feature is removed from the target cost, then, the performance of the target cost function should deteriorate. The extend to which it deteriorates when a target feature is excluded, quantifies the feature's importance. We estimate the relative importance of a target feature based on the deterioration of selection performance when a target feature is excluded from the target cost. This is explained in detail in the following discussion. For simplicity of notation, we stop showing a candidate and a target with the target cost function. Lets assume that the current set of target features is $\mathbb{F}$, and current feature being considered is f. Lets denote the singleton feature set $\{f\}$ with F, $F^c = \mathbb{F} - F$. Let us express the target cost function as follows:

$$TC = \underbrace{w_F TC_F + (1 - w_F)\overbrace{TC_{F^c}}^{(2)}}_{(1)} \tag{5.13}$$

The target cost (TC) shown above is the weighted sum of the following two components:

(a) The target cost function with one feature f, TC_F.

(b) A target cost function which excluded feature f, from the target feature set, TC_{F^c}.

The target cost function highlighted as (1) in the above equation takes all the features into account and the target cost function highlighted as (2) above excludes feature f. Using (1) and (2) as the two target costs, two disorders are calculated. The disorder calculated using (1) is referred to as Combined Disorder (CD), which depends also on w_F. The disorder calculated using (2) is refereed to as Exclusion Disorder (ED). The following can be said with respect to the comparison of CD and ED:

- A feature f is considered to contribute information, if disorder increases when its excluded from the target cost: $ED_f > CD$.

- A feature f is considered to contribute noise, if the disorder decreases when its excluded from the target cost: $ED_f < CD$.

Those features which contribute information, their weights should be increased proportional to their contribution, features which seem to contribute noise, their weights should be decreased till they become contributing features; if a feature contributes only noise (for long), they are eliminated from the feature set.

The following possibilities need to be considered while classifying features informative:

(1) Features might provide information if given an optimum weight (in the weigh combination). Excluding these features might modify the disorder compared to their inclusion and the increase or decrease depends on the combination of weights.

(2) Features which don't provide any information will not affect the disorder with their exclusion and inclusion even with a change in their relative weight in the target cost.

(3) Features which contribute only noise by their inclusion in the target cost, regardless of the non-zero weight given to them, the combined disorder will always be greater than the disorder with their exclusion.

Based on this analysis we developed an iterative algorithm. At any iteration, the weights are updated based on the comparison of ED of different features and CD as follows:

- Those features for which $ED > CD$, their weights increase. The increase in proportional to the difference in ED and CD.

- Those features for which $ED < CD$, they can belong to either category (1) or (3). The feature weights are updated proportional to the difference in CD and ED. A feature which shows this trend ($ED < CD$) for long, it is eliminated from the feature set.

- Features belonging to category (2) are also eliminated ($ED = CD$).

- A fraction of total weight from the set of features for which ($ED < CD$) is distributed among features for which ($ED > CD$).

- To make the change in the weights slow, the weights at each iteration are made a function of the previous iteration. Any new weight after an iteration, is a fraction (fixed parameter) of the old weight and the change based on the difference in CD and its ED.

5.3.2.4 Algorithm

We provide the precise details of the algorithm here. **Notation:** For any iteration i, the complete set of features is $\mathbb{F}_i$; a singleton set having feature f is denoted by the set F; the set of features excluding a feature f from set $\mathbb{F}$ is $F_i^c = \mathbb{F}_i - F$; the disorder with the complete set of features and their weights at iteration i (from previous iteration) i.e., the combined disorder CD is $\Delta(i)$; the disorder with a feature f excluded from the target cost (ED) is $\Delta_{F_i^c}(i)$; set of all the features for which $\Delta_{F_i^c}(i) > \Delta(i)$ is denoted by $\mathbb{F}_i^+$ and $\mathbb{F}_i^-$ for those which are qualified to remain in the feature set with $\Delta_{F_i^c}(i) < \Delta(i)$; set of all features which are being eliminated are $\mathbb{F}_i^0$. For a feature f, $t_f(i)$ is the number of iterations it has been in $\mathbb{F}_i^-$ consecutively till iteration i without being eliminated.

At every iteration i the following quantities are calculated for updating the feature weights:
Information Component ($I_F(i)$): For a feature $f \in \mathbb{F}_i^+$, i.e. $\Delta(i) < \Delta_{F_i^c}(i)$:

$$I_F(i) = \frac{|\Delta(i) - \Delta_{F_i^c}(i)|}{\sum\limits_{a \in \mathbb{F}_i^+} (|\Delta(i) - \Delta_{A_i^c}(i)|)} \tag{5.14}$$

Noise Component $N_F(i)$: For a feature $f \in \mathbb{F}_i^-$ and $\Delta(i) > \Delta_{F_i^c}(i)$:

$$N_F(i) = \frac{\Delta(i) - \Delta_{F_i^c}(i)}{\sum\limits_{a \in \mathbb{F}_{-i}} (\Delta(i) - \Delta_{A_i^c}(i))} \tag{5.15}$$

Based on this $N_F^{'}(i)$ calculated as follows to update the weight at every iteration.

$$N_F^{'}(i) = \frac{(1 - N_F(i))}{(n_{\mathbb{F}_i^-} - 1)} \tag{5.16}$$

where, $n_{\mathbb{F}_i^-}$ is the number of elements in the set $\mathbb{F}_i^-$. $N_F^{'}(i)$ increases as $N_F(i)$ decreases, so features which contribute more noise will lose more weight in the target functions subsequently. In case there is only one feature in $\mathbb{F}_i^-$, then $N_F^{'}(i) = 1$.

The following are the parameters of algorithm:

- T, the maximum number of tolerant iterations for a noisy feature. A feature f for which $t_f(i) > T$ is eliminated from the feature list. If a feature f changes from set $\mathbb{F}_i^-$ to set $\mathbb{F}_i^+$ in an iteration i, then $t_F(i)$ is set to 0.

- α_- and α_+, the fractions of weights of any features in $\mathbb{F}_i^-$ and $\mathbb{F}_i^+$ respectively that is

carried forward from the weight in the previous iteration. This makes the updated weight in the current iteration a function of the weight in the previous iteration. It is done to make the change in weights slow.

- β is the fraction of the total changeable weight in $\mathbb{F}_i^-$ that is gained by features in $\mathbb{F}_i^+$. The logic behind this distribution is that, features in $\mathbb{F}_i^-$ loose weight while features in $\mathbb{F}_i^+$ gain weight.

- Maximum allowed iterations, for which the algorithm is executed. This is fixed based on the rate of change in total disorder (decrease in combined disorder per iteration).

The goal of the algorithm is to select the set of features and tune their respective weights in such a way that the *disorder* Δ described by equation (5.8) is minimized:

- **Beginning:** Target cost function with the complete set of features which are assigned equal weights.

- **At every iteration i:**

  ⋆ The following are first determined:

    ◦ $\Delta(i)$.

    ◦ for all $f \in \mathbb{F}_i$: $\Delta_{F^c}(i)$.

  ⋆ Elimination of all those features f for which one of the following conditions is satisfied:

  1. $(\Delta(i) - \Delta_{F^c}(i)) \approx 0$
  2. $(\Delta(i) - \Delta_{F^c}(i)) > 0$ and $t_F(i) > T$

  ⋆ Update weights: The update is such that the change is slow. For that, a fraction of weight (α_+ for features in $\mathbb{F}_i^+$ and α_- for features in $\mathbb{F}_i^-$) remains constant with respect to the previous iteration.

    ◦ For a feature $f \in \mathbb{F}_i^+$: More the information in the feature, higher the weight.

$$\begin{aligned} w_F(i) = \quad & \alpha_+ w_F(i-1) \quad (1) \\ & + \\ & W_{\mathbb{F}_i^+} I_F(i) \quad (2) \end{aligned} \tag{5.17}$$

The first component (1), depends on the feature weight in the previous iteration; the second component (2), depends on the information component of the feature. $W_{\mathbb{F}_i^+}$ is the total weight that will be redistributed in $\mathbb{F}_i^+$. $W_{\mathbb{F}_i^+}$ is calculated as

follows:

$$\begin{aligned} W_{\mathbb{F}_i^+} = \quad & (1-\alpha_+) \sum_{a \in \mathbb{F}_i^+} w_A(i-1) \quad (i) \\ & + \\ & (1-\alpha_-)\beta \sum_{b \in \mathbb{F}_i^-} w_B(i-1) \quad (ii) \\ & + \\ & \sum_{c \in \mathbb{F}_i^0} w_C(i-1) \quad (iii) \end{aligned} \tag{5.18}$$

The first component (i), is the total changeable weight of features in $\mathbb{F}_i^+$; the second component (ii), is the total changeable weight of features in $\mathbb{F}_i^-$ that is gained by features in $\mathbb{F}_i^+$; the third component (iii), is the total weight of the features being eliminated, $\mathbb{F}_i^0$. The total weight of the features being eliminated $\mathbb{F}_i^0$ is re-distributed among features in $\mathbb{F}_i^+$.

- For a feature $f \in \mathbb{F}_i^-$: Lesser the noise contribution, higher the weight.

$$\begin{aligned} w_F(i) = \quad & \alpha_- w_F(i-1) \quad (1) \\ & + \\ & W_{\mathbb{F}_i^-} N'_F(i) \quad (2) \end{aligned} \tag{5.19}$$

The first component (1), depends on the weight of the feature f in the previous iteration; the second component (2), depends on the Noise Component of feature f. $W_{\mathbb{F}_i^-}$ is the fraction of total changeable weight of features in $\mathbb{F}_i^-$ that is redistributed to features in $\mathbb{F}_i^-$ itself. It is calculated as follows:

$$W_{\mathbb{F}_i^-} = (1-\alpha_-)(1-\beta) \sum_{a \in \mathbb{F}_i^-} w_A(i-1) \tag{5.20}$$

- **Termination:** The algorithm is terminated when maximum number of allowed iterations are executed or when there is no improvement (decrease in combined disorder) in an iteration beyond a certain ϵ. The best weights w.r.t the least disorder along all the iterations are chosen for the final target cost for the phoneme.

5.3.3 Application to AV target cost function tuning

The visual speech features vectors x_u, x_v of equation (5.11) were the first 12 PCA coefficients. For acoustic speech, MFCC and f0 were used, where the 13 MFCC were extracted at the rate of 100Hz and f0 extracted every 8 milliseconds respectively.

The parameters of the algorithm were chosen based on the trade-off between time required for each iteration, speed of change of disorder which affects the required minimum number of iterations for the attainment of relative convergence (when the rate of change of disorder is low) . By trial and error on a single phoneme /a/ which has a good coverage, the following parameters were finally chosen for the weight tuning of all the phonemes:

- T, the maximum tolerant iterations is 2. A feature f is removed whenever its $t_F(i) > 2$.
- $\alpha_+ = 0.5$ and $\alpha_- = 0.5$; $\beta = 0.05$.

The tuning was done for 5 weight combinations. The result of selected target features with only one of the dissimilarity measures (duration, visual, MFCC and f0) i.e., only one of the following $\{w_{dur}, w_{ac}, w_{vs}, w_{f0}\}$ being one and all others 0 was analyzed and all the measures taking equal weights. The first four weight combinations were chosen for the analysis of target features with respect to each of these necessary aspects. The fifth weight combination is chosen for the final weight tuning to be used in the system for selection. This weight combination $(0.25, 0.25, 0.25, 0.25)$ performed reasonably well with respect to informal listening tests. This can be further improved based on the analysis of perceptual evaluation and correlation with the objective evaluation. For each of the weight combinations, this algorithm has been executed separately for all the phonemes in the phoneme set using our corpus to obtain different target functions, i.e., different set of features and their weights for different phonemes.

5.3.4 Analysis of selected features and their relative importance

In this section, we present the analysis of target features based on their relative importance for each of the constituent aspects included in the dissimilarity metric: pitch, local acoustic speech, duration and visual speech. They are based on target feature weighting by taking one constituent metric at a time in the dissimilarity metric. The features with lower weights (< 0.01) are not shown in this analysis. These results are presented for vowels and consonants separately. Linguistic features can describe a current candidate or its left or right context. Phonetic features can describe a candidate's left or right context (see section 5.1). To analyze the results, we calculate the mean and standard deviation of weights assigned to each feature by taking together the context and the current candidate. The weights are assigned such that the sum of the weights over all the target features is 1. These results are shown in tables 5.5 to 5.12.

- **Pitch:** For vowels, mean total weight given to linguistic features is 0.19 and 0.81 to phonetic features with a standard deviation of 0.24. For consonants, linguistic features get

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Voicing	0.71	0.28	Voicing	0.26	0.32
Kind	0.08	0.13	Kind	0.13	0.15
			Lip shape	0.13	0.19
			Manner of articulation	0.11	0.14

Table 5.5: Phonetic features important for pitch

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Left silence	0.05	0.07	Right silence	0.14	0.21
Syllable position in RG	0.04	0.06	Syllable position in RG	0.07	0.08
Word position in sentence	0.03	0.07	Syllable position in word	0.04	0.07
Phoneme number in syllable	0.03	0.09	Word position in RG	0.03	0.05
Right silence	0.02	0.03	Word position in sentence	0.02	0.06
Syllable position in word	0.01	0.01	Phoneme number in syllable	0.02	0.02
			Syllable number in sentence	0.01	0.01
			Syllable kind	0.01	0.01
			Word number in RG	0.01	0.03

Table 5.6: Linguistic features important for pitch

0.36 as the mean total weight and 0.64 for phonetic features with a standard deviation of 0.26. The preceding context is important in terms of both phonetic and linguistic features for pitch prediction. The list of important linguistic and phonetic features with the mean and standard deviation of weights for vowels and consonants is given in tables 5.5 and 5.6.

- Phonetic features: For both vowels and consonants, contextual phoneme voicing and phoneme kind are important features. For consonants, lip shape during articulation and manner of articulation are also important.
- Linguistic features: For both vowels and consonants, relative position of nearest following and preceding silence, syllable position in rhythm group(RG) and word, phoneme number in a syllable and word position in a sentence are important.

- **Local speech acoustics:** The acoustic features considered (MFCCs) can be assumed to describe local speech acoustics. For vowels, phonetic features get total mean weight of 0.67 and 0.33 for linguistic features with a standard deviation of 0.26. For consonants, the total mean weight for linguistic features is 0.19 and 0.81 for phonetic features, with a standard deviation of 0.12. The list of important linguistic and phonetic features with the mean and standard deviation of weights for vowels and consonants is given in tables 5.7 and 5.8.

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Voicing	0.26	0.25	Lip shape	0.32	0.20
Place of articulation	0.21	0.22	Place of articulation	0.20	0.27
Manner of articulation	0.13	0.11	Voicing	0.12	0.19
Kind	0.04	0.06	Manner of articulation	0.10	0.12
Lip shape	0.03	0.04	Kind	0.07	0.10

Table 5.7: Phonetic features important for local speech acoustics

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Syllable position in word	0.07	0.11	Syllable position in RG	0.04	0.05
Left silence	0.05	0.05	Right silence	0.04	0.06
Syllable position in RG	0.04	0.04	Left silence	0.02	0.02
Word position in sentence	0.04	0.06	Syllable position in word	0.02	0.02
Phoneme number in syllable	0.04	0.06	Word position in RG	0.01	0.01
Syllable kind	0.04	0.05	Phoneme number in syllable	0.01	0.01
Right silence	0.02	0.03	Word position in sentence	0.01	0.01

Table 5.8: Linguistic features important for local speech acoustics

  - Phonetic features: For vowels, voicing of the preceding phonemes, place and manner of articulation of the following phoneme are the most important features, followed by place of articulation of the preceding and voicing of the following phoneme. For consonants, lip shape of the following phonemes seems to be the most important feature besides place of articulation and kind of the following phonemes. Just as in the case of f0, voicing of the preceding phoneme is also an important feature.
  - Linguistic features: For both vowels and consonants, syllable position in word and RG, relative position of the nearest left and right silence, phoneme number in a syllable, word position in a sentence are important. Syllable kind and word position in sentence are also important for vowels and consonants respectively.

- **Duration:** For duration, linguistic features are dominant and invariably the most important compared to phonetic features. The pattern is even more pronounced in the case of vowels. For vowels and consonants, the total mean weight assigned to linguistic features is 0.65 and 0.62 respectively, and the standard deviation is 0.18 and 0.25 respectively. The list of important linguistic and phonetic features with the mean and standard deviation of weights for vowels and consonants is given in tables 5.9 and 5.10.

  - Phonetic features: For both vowels and consonants kind of following phoneme is the

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Kind	0.25	0.16	Kind	0.15	0.13
Lip shape	0.05	0.06	Manner of articulation	0.10	0.16
Place of articulation	0.03	0.08	Voicing	0.08	0.13
Manner of articulation	0.02	0.04	Lip shape	0.04	0.09
Voicing	0.01	0.02	Place of articulation	0.02	0.02

Table 5.9: Phonetic features important for duration

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Syllable position RG	0.41	0.22	Syllable position in RG	0.23	0.15
Syllable position in word	0.07	0.09	Right silence	0.16	0.23
Right silence	0.04	0.09	Left silence	0.08	0.12
Syllable kind	0.04	0.08	Syllable position in word	0.03	0.05
Left silence	0.02	0.04	Word position in RG	0.03	0.08
Phoneme number in syllable	0.02	0.04	Phoneme number in syllable	0.02	0.05
Word position in RG	0.02	0.03	Syllable number in RG	0.02	0.02
			Word number in RG	0.01	0.02
			Syllable number in sentence	0.01	0.02

Table 5.10: Linguistic features important for duration

most important feature. For consonants, the manner of articulation and voicing of the following contextual phoneme is also important.

- Linguistic features: For both vowels and consonants, the syllable position in the RG is the most important feature, followed by relative positions of left and right silence, syllable position in word, phoneme number in a syllable, word position in a RG.

- **Visual features:** For visual speech, the total mean weight assigned to linguistic features is 0.31 for vowels and 0.12 for consonants with a standard deviation of 0.17 and 0.10 respectively. The list of important linguistic and phonetic features with the mean and standard deviation of weights for vowels and consonants is given in tables 5.11 and 5.12.

  - Phonetic features: For vowels, place of articulation of the following and preceding phonemes are the most important features in the decreasing order of importance. The lip shape during articulation and manner of articulation of the contextual phonemes are also observed to be important. For consonants, lip shape of the following phoneme, lip shape of the preceding phoneme and place of articulation of the preceding phoneme are observed to be the 3 most important features in the decreasing order of importance.

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Place of articulation	0.36	0.18	Lip shape	0.77	0.16
Lip shape	0.14	0.19	Place of articulation	0.04	0.05
Manner of articulation	0.09	0.09	Voicing	0.02	0.03
Voicing	0.07	0.09			
Kind	0.04	0.06			

Table 5.11: Phonetic features important for visual speech

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Syllable position in word	0.11	0.11	Syllable position in word	0.03	0.05
Syllable kind	0.04	0.07	Syllable number in sentence	0.02	0.02
Syllable number in Sen	0.04	0.02	Right silence	0.01	0.02
Phoneme number in syllable	0.02	0.03	Word position in sentence	0.01	0.02
Right silence	0.02	0.04			
Word position in sentence	0.02	0.01			
Word number in RG	0.02	0.05			

Table 5.12: Linguistic features important for visual speech

- Linguistic features: For vowels, syllable position in a word is an important feature.

The analysis of these selected features is in itself an interesting problem. The relative importance of the contextual features indicates that the right context is more important than the left. This is more pronounced in phonetic features weights. One of the possible interpretations of this is that the instances of anticipatory coarticulation is higher than the instances of carryover coarticulation in French. Word number in sentence has got eliminated for most of the phonemes as the corpus is not sufficient to establish any such relation. Numeric features in general have got lower weights which show that the relative position is more important than their exact position. The former features are size invariant. For example, 'syllable position in RG' does not depend on the total number of syllables in RG. But 'syllable number in RG' depends on the total number of syllables in RG. The selected features and their relative weights implicitly indicate the validity of the algorithm. For example, for pitch and duration, syllable position in RG, relative position of nearest left and right silence, syllable postion in word are shown to be important. These features are known to be important for explaining many of the prosodic patterns in French.

With the fifth combination with equal weights to all the four constituents of the dissimilarity metric, the selected features contain the features which are important for all the four constituent aspects (see tables 5.13 and 5.14). The total mean weight for linguistic features in case of vowels

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Voicing	0.48	0.27	Lip shape	0.35	0.15
Kind	0.13	0.16	Voicing	0.17	0.22
Place of articulation	0.06	0.05	Place of articulation	0.10	0.10
Manner of articulation	0.03	0.02	Kind	0.08	0.10
Lip shape	0.02	0.02	Manner of articulation	0.04	0.04

Table 5.13: Phonetic features for acoustic-visual speech

Vowels			Consonants		
	Weight			Weight	
Feature	μ	σ	Feature	μ	σ
Syllable position in RG	0.09	0.08	Right silence	0.10	0.13
Right silence	0.04	0.05	Syllable position in RG	0.06	0.07
Left silence	0.04	0.06	Syllable position in word	0.02	0.02
Syllable position in word	0.04	0.07	Left silence	0.01	0.02
Phoneme number in syllable	0.03	0.05	Ph number in syllable	0.01	0.03
Syllable Kind	0.02	0.03	Word position in sentence	0.01	0.01
Word position in sentence	0.01	0.01			

Table 5.14: Linguistic features for acoustic-visual speech

and consonants are 0.28 and 0.26 respectively, and the standard deviation is 0.24 and 0.17 respectively. We use these features and their weights determined in our synthesis system. We present the objective and perceptual evaluation done for the synthesized speech using these feature weights.

5.4 Conclusion

In this chapter, we have presented the set of corpus-independent target features and explained the corpus-based visual target features that we developed for improving synthesis with our corpus. We detailed the iterative target feature weighting technique that we have designed. It assigns weights and performs elimination of redundant features simultaneously. We finally presented the analysis of the patterns that were observed in the selected features and their weights. The relative weighting of the target feature affects selection and hence the synthesis results. Majority of the observations with respect to selected features and their relative weights are in agreement with the phonetic and linguistic studies which show the strength of this algorithm. It also indicates that the constituent metrics included to represent pitch, duration, local speech acoustics and visual speech are indeed correlated to these aspects.

The weight tuning algorithm that we presented (section 5.3.2) performs automatic weight tuning based on (1) a dissimilarity metric which compared the difference in two speech re-

alisations and (2) a set of target features used to describe the targets and candidates. The performance of selection based on the resultant target cost depends on various factors. Firstly, For the various aspects included, different distance measures can be investigated with respect to their correlation with human perception. Such studies have been done with respect to acoustic concatenation costs (Wouters and Macon, 1998; Vepa et al., 2002; Klabbers and Veldhuis, 1998). Secondly, the importance of the different aspects of dissimilarity metric varies among phonemes. For example, it is known that vowel durations are more important for good prosody. The two above mentioned factors require substantial investigation. Lastly, the weights given to these constituent metrics might further improve by systematic and extensive perceptual experiments with human participants. It can be argued that this process is inefficient and slow. But, a good justification to such an approach is that weight tuning problem in the huge dimensional space of target features is being mitigated by setting the weights of constituents of the dissimilarity metric which is a much smaller dimension. Also, since the synthesized speech is targeted for humans, reinforcement from human participants is advantageous. We performed evaluations through human subjects to assess the final system with the resultant target features and their weights using the weight tuning algorithm. In the following chapter, we describe these tests besides summarizing objective evaluation techniques that we have used while developing selection strategies[3].

[3] A part of this chapter was published in (Musti et al., 2011).

Chapter 6

Evaluation

Throughout the development process, the different methodologies being used to improve synthesis were systematically and automatically evaluated. This objective evaluation was based on some metrics that we defined. This evaluation can be performed either by comparing synthesized AV speech signals to real speech signals, or based on a comparison with corpus statistics. However, as this acoustic-visual speech synthesis system is targeted for humans, the system should be evaluated using perceptual experiments where human beings are the center of this evaluation. In the context of audio-visual speech, the evaluation of both the channels is not straightforward and requires a careful consideration of the various factors which might affect the synthesis quality and the limitations of the system while setting benchmarks for comparison.

In this chapter, we first describe the various objective evaluation metrics used for evaluating different selection techniques (in section 6.1). In section 6.2, we describe the perceptual and subjective evaluations done along with their results. Finally, we present a preliminary analysis of the subjective evaluation results in comparison with the objective evaluation metrics in section 6.3. [4]

6.1 Objective evaluation

For a fast automatic evaluation of the synthesized speech, it is a general practice to leave some of the sentences outside the synthesis corpus for comparison purpose. They are generally either specially designed or chosen based on some necessary conditions. They are considered as references for comparative evaluation. We have a set of 20 test sentences which are not part of the synthesis corpus for comparative evaluation.

[4] A short overview of our system and evaluation results presented in this chapter were published in (Musti et al., 2012)

6.1.1 Objective evaluation based on comparison of two signals

We have utilized three objective evaluation metrics which have been introduced in the previous chapter (section 5.3) and the correlation coefficient and root mean squared error (RMSE) between real and synthesized test sentences. To make the duration (number of samples) equal in both sentences a simple linear interpolation is applied for each demi-phones wherever necessary (see Fig. 5.5). Lets assume that, x_d and y_d are the sequences of the d^{th} acoustic or visual parameters of a real and synthesized sentence respectively having n samples. Then, the first two metrics are calculated as follows:

- Pearson's Correlation Coefficient: the correlation coefficient $r_{x_d y_d}$ is calculated as follows:

$$r_{x_d y_d} = \frac{n \sum_i x_d(i) y_d(i) - \sum_i x_d(i) \sum_i y_d(i)}{\sqrt{n \sum_i x_d(i)^2 - (\sum_i x_d(i))^2} \sqrt{n \sum_i y_d(i)^2 - (\sum_i y_d(i))^2}} \tag{6.1}$$

- Root Mean Squared Error (RMSE) $d^{rmse}(x_d, y_d)$ is calculated as follows:

$$d^{rmse}(x_d, y_d) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_d(i) - y_d(i))^2} \tag{6.2}$$

The considered acoustic parameters were the first 13 MFCCs and F0. The considered visual parameters were the first 12 PCA coefficients.

The duration based metrics are calculated as follows:

1. For the purpose of comparing any two candidates u and v of the same phonetic label for the purpose of target weight tuning the following metric was used:

$$D^{dur}(u, v) = \frac{|(dur_u - dur_v)| - dur_{min}(p)}{dur_{max}(p) - dur_{min}(p)} \tag{6.3}$$

   Where, $dur_{max}(p)$ and $dur_{min}(p)$ are the maximum and minimum of durations of all the candidates for phoneme p; and dur_u and dur_v are the durations of candidate u and v.

2. For the purpose of comparing two whole sentences (segment wise), the following duration metric was used:

$$D^{dur}(s, r) = \frac{\sum_{j=0}^{N} \frac{|(dur_s(j) - dur_r(j))|}{dur_r(j)}}{N} \tag{6.4}$$

   Where, s and r are the synthesized and real sentences respectively having N phonetic segments and $dur_s(j)$ and $dur_r(j)$ are the durations of j^{th} phonetic segments of real and synthesized sentences.

6.1.2 Objective evaluation based on statistical analysis and thresholds

Sometimes objective evaluation metrics which are based on statistical analysis of the corpus are developed and utilized for various purposes. For the purpose of total cost weight tuning for setting the weights of the target cost, acoustic and visual join costs, we utilized three objective evaluation metrics which belong to this category. We first calculated the standard deviation of the first PCA coefficient (denoted by σ_{PC1}) from the whole corpus. Similarly, standard deviation of its first order derivative (denoted by $\sigma_{\Delta PC1)}$) from the whole corpus was also calculated. Then, for a set of synthesized sentences, the sentences were scanned at all the concatenation boundaries to count the following:

- Total instances where the differences between first PCA coefficients exceed ϵ_{pc1}.
- Total instances where the differences between first order derivative of first PCA coefficients exceed $\epsilon_{\Delta PC1}$.
- Total instances where the differences in f0 exceed ϵ_{f0}. Bark was chosen as the suitable perceptual scale.

The first principal component and its derivative were chosen as the first principal component itself accounts for about 57% of the data variance and also gives an indication of the discontinuity in the subsequent components. These values give an indication of the visual and acoustic discontinuation at the concatenation boundaries. These values along with a duration were used for evaluating the efficiency of the total cost function. Besides the above 3 metrics, a duration metric based on the comparison of real and synthesized sentences was also used as explained below.

- Total instances of vowels where the difference in duration ratio of synthesized and real sentences is greater than ϵ_{dur}.

The thresholds were chosen empirically by perceptual experimentation. In this case the considered thresholds were $\epsilon_{pc1} = 0.5\sigma_{PC1}$, $\epsilon_{\Delta PC1} = 0.5\sigma_{\Delta PC1}$, $\epsilon_{f0} = 0.25$Barks and $\epsilon_{dur} = 150\%$. Throughout the development process, this method was applied for the tuning of the total cost weights, whenever we modified the components of target cost function or concatenation cost function. The following weights were used for the total cost function for selection, $w = 1$, $w_{aj} = 0.943$ and $w_{vj} = 0.897$, where w, w_{aj} and w_{vj} are the weights assigned to target, acoustic concatenation and visual concatenation cost functions respectively. [5]

[5] This work was mainly done by Asterios TOUTIOS and was presented in (Toutios et al., 2011).

6.2 Human-centered evaluation

To evaluate our overall audio-visual speech synthesis system, the following perceptual intelligibility and subjective quality evaluation tests were conducted. In these tests a total of 39 participants between 19 to 65 years of age with normal auditory and visual abilities participated. Among the participants, 15 were female and the rest were male participants. All these participants were native French speakers. The tests were conducted across internet where each of the participants heard and saw the stimuli in their computers with the available hardware. A set of basic instructions was played at the beginning of these tests.

6.2.1 Intelligibility tests

The intelligibility test was at the word level. Each human subject was presented with 50 one or two syllabic French words and asked to recognize and report the word. Some examples of the words that were presented include { anneau (ring), bien (good), chance (luck), pince (clip), laine (wool), cuisine(kitchen) }. Among these words, 11 were those which are present in the corpus. These in-corpus words were included to set a benchmark for the best possible intelligibility by the recorded data.

These tests were done at two levels: (1) acoustic-only speech, (2) audio-visual speech. In each of these categories, the acoustic speech component was degraded to two noise levels. Hence, each word was played 4 times: (1) acoustic-only with low noise component (SNR of -6 dB), (2) acoustic-only with high noise component (SNR of -10 dB), (3) audio-visual with low noise (SNR -6dB), (4) audio-visual speech with high noise (SNR of -10 dB). The addition of noise also ensures that the listener pays attention to the visual modality of speech. The aim is to evaluate both visual and acoustic modalities, and also to estimate the advantage of audio-visual speech over acoustic-only speech. These noise thresholds were decided based on the several audio-visual perceptual experiments to strike a trade-off between these two objective. The facial animation is shown as the 3D surface of the face using sparse mesh, which has the dynamics of facial deformations, but without the texture and color information (see Fig. 3.9). Besides, the information regarding internal articulators, teeth and tongue is also missing from the animations.

Table 6.2 includes the intelligibility scores in terms of the fraction of the total words recognized in each of the four categories by different users. Table 6.1 shows the mean intelligibility scores of in-corpus words and out-of-corpus words. Any word completely recognized correctly is classified as a correct response. The intelligibility results of the in-corpus words shows the best possible results with the corpus we have recorded. These in-corpus intelligibility results show that the best possible intelligibility with our corpus is not very high. The comparatively

	Audio		Audio-Visual	
	Low Noise	High Noise	Low Noise	High Noise
In-Corpus words	0.69	0.59	0.72	0.65
Out-of-Corpus words	0.40	0.34	0.45	0.40

Table 6.1: Mean intelligibility scores

lower results for the in-corpus words can be attributed to the absence of internal articulators. The difference in performance between in-corpus and out-of-corpus words in the acoustic domain show the possibility of further improvement. Results show that the addition of visual component to the acoustics improves intelligibility. The intelligibility in noisy environment is an important aspect to evaluate AV speech. The intelligibility results only confirm this. This is also interesting because, visual speech rendering though far from being photo-realistic is still effective in presenting the articulatory dynamics. Another general observation that is confirmed by these results is that the improvement in speech recognition is more in high-noise to low-noise speech acoustics. The advantage of the addition of visual speech is more obvious in the out-of-corpus words. These results are interesting also because in spite of the internal articulators being absent from the animations, the results show the advantage of AV speech over acoustic speech. This shows that the visual and acoustic speech are in agreement to each other.

6.2.2 Quality evaluation tests

Subjective tests were performed for the evaluation of the synthesis quality. 20 audio-visual sentences were played, out of which 7 sentences were real and the rest (13 sentences) were synthesized sentences which correspond to a subset of the test sentences we have for objective evaluation purpose. Just as in the case of intelligibility tests, the five real sentences serve as the best response that is possible with the corpus utilized for synthesis which affects various aspects of the synthesized speech like duration, phonetic coverage and facial speech rendering technique. For each of the stimulus, 5 questions were posed and participants were asked to give categorical responses based on the 5 point MOS scale. These 5 questions and the possible categorical answers are given in table 6.3. The first question (Q1) represents the synchrony in the acoustic and visual modalities. The second question (Q2) implicitly represents the prosody. Third and fourth questions (Q3 and Q4) are representative of the naturalness of acoustic and visual modalities respectively. The last question (Q4) is representative of the overall speech quality and pleasantness. The subjective evaluation results for in-corpus and out-of-corpus sentences are given in table 6.4. The results to the question Q1 show that the audio-visual alignment is good, and the acoustic prosody is acceptable (Q2 results). It has to be highlighted

Participants	Audio		Audio-Visual	
	Low N.	High N.	Low N.	High N.
1	0.48	0.46	0.56	0.46
2	0.26	0.28	0.36	0.32
3	0.46	0.32	0.44	0.52
4	0.56	0.44	0.52	0.52
5	0.44	0.30	0.56	0.44
6	0.54	0.52	0.54	0.44
7	0.42	0.26	0.44	0.36
8	0.50	0.42	0.52	0.50
9	0.38	0.24	0.44	0.38
10	0.36	0.28	0.44	0.32
11	0.52	0.44	0.58	0.46
12	0.46	0.44	0.50	0.42
13	0.52	0.30	0.54	0.42
14	0.34	0.26	0.40	0.24
15	0.50	0.42	0.46	0.42
16	0.40	0.28	0.48	0.40
17	0.54	0.46	0.60	0.58
18	0.48	0.46	0.54	0.50
19	0.52	0.50	0.58	0.56
20	0.46	0.42	0.56	0.52
21	0.40	0.42	0.42	0.38
22	0.44	0.44	0.54	0.50
23	0.52	0.42	0.58	0.54
24	0.68	0.62	0.76	0.70
25	0.56	0.40	0.72	0.64
26	0.56	0.32	0.48	0.50
27	0.58	0.54	0.62	0.56
28	0.40	0.34	0.42	0.46
29	0.44	0.40	0.52	0.44
30	0.50	0.40	0.56	0.46
31	0.36	0.30	0.42	0.32
32	0.48	0.42	0.46	0.42
33	0.48	0.34	0.46	0.46
34	0.40	0.36	0.42	0.36
35	0.40	0.36	0.40	0.28
36	0.44	0.40	0.44	0.42
37	0.38	0.38	0.50	0.50
38	0.38	0.44	0.46	0.40
39	0.62	0.62	0.60	0.60
Mean	0.47	0.40	0.51	0.46
Std dev.	0.08	0.09	0.09	0.10

Table 6.2: Intelligibility Results in the four categories, *acoustic-only + high noise*, *acoustic-only + low noise*, *audio-visual + high noise* and *audio-visual + low noise*

	Question	Categorical responses
Q1.	Does the lip movement match the pronounced audio?	(5) Always – (1) Never
Q2.	Is this sentence an affirmation (neutral reading)?	(5) Totally agree – (1) Not at all
Q3.	Is the acoustic speech natural?	(5) Very natural – (1) Not natural
Q4.	Is the facial animation natural?	(5) Very natural – (1) Not natural
Q5.	Is the pronunciation of this sentence by the talking head pleasant?	(5) Very pleasant – (1) Not at all

Table 6.3: This table shows the five questions and the expected categorical responses for evaluating the quality of the synthesized speech

	Question-1	Question-2	Question-3	Question-4	Question-5
Overall	3.88	3.93	3.04	2.92	3.02
Out-of-Corpus sentences	3.76	3.78	2.57	2.80	2.65
In-Corpus sentences	4.80	4.91	4.56	3.67	4.32

Table 6.4: Mean MOS scores for the five questions

that the prosody was generated without using any explicit model. The naturalness scores for voice seem to be low as shown in the Q3 results. These can be attributed to the relatively small size of the corpus and consequently the absence of some diphones in the corpus. On the contrary, the naturalness scores of facial animation (Q4 results) are high. This shows that articulatory dynamics are being represented well. Further, there might be a small component of the fact that the facial representation or 'human likeness' is not close to the uncanny valley and so participants are not very critical.

6.3 Analysis of perceptual evaluation for better objective metrics

The objective evaluation metrics calculated for the out-of-corpus sentences on the whole sentences are given in table 6.7. These results in comparison with those given in table 6.6 show

	Question-1	Question-2	Question-3	Question-4	Question-5
1	4.38	4.25	3.72	3.42	3.70
2	3.92	4.43	3.60	3.08	3.50
3	4.12	4.43	4.12	3.22	4.12
4	3.75	4.00	4.03	2.97	3.72
5	4.15	4.28	3.92	3.10	3.53
6	3.97	3.62	3.80	2.97	3.40
7	4.38	4.32	3.97	3.25	3.83

Table 6.5: Mean MOS scores for the five questions asked to evaluate the quality of the audio-visual speech synthesis for each of the in-corpus sentences

	Question-1	Question-2	Question-3	Question-4	Question-5
1	3.78	3.70	2.45	2.50	2.53
2	3.85	4.25	3.03	3.00	3.08
3	3.42	3.85	2.53	2.78	2.55
4	3.78	3.67	2.58	2.78	2.60
5	3.65	3.15	2.30	2.60	2.40
6	4.05	3.75	2.60	2.85	2.62
7	3.12	3.20	2.03	2.50	2.17
8	4.15	4.40	3.30	3.17	3.30
9	3.70	3.92	2.67	2.88	2.62
10	3.38	3.55	2.12	2.78	2.30
11	4.20	3.58	2.00	2.72	2.25
12	3.53	3.95	2.42	2.83	2.75
13	4.15	4.10	3.35	3.17	3.40

Table 6.6: Mean MOS scores for the five questions asked to evaluate the quality of the audio-visual speech synthesis system for out-of-corpus sentences

that the correlation of the two are not very high on a per-sentence basis. To investigate for the perceptually important segments which affect these subjective evaluation results, they were analyzed in comparison with the objective evaluation metrics explained in section 6.1. The analysis was based on the acoustic and visual modality. For this purpose different phoneme sets belonging to different categories were considered; like, all-phonemes, vowels, consonants, voiced phonemes, unvoiced phonemes, visible phonemes, visible vowels, not-visible phonemes etc. Visible phonemes are those which have identifiably unique visible articulation, like /p/, /o/ etc. The visible phoneme set includes those phonemes which are shown to have good recognition based on visual features (chapter 4). The out-of-corpus sentences are a subset of the test sentences for which we have the real utterances, i.e. real acoustic and visual speech realization. For each out-of-corpus sentence, the objective evaluation metrics were calculated by comparing the synthesized and real utterances as follows:

- For each phoneme category, overall objective evaluation metrics mentioned were calculated. For example, considering only vowel segments, for each sentence the overall objective evaluation metrics are calculated. We refer to these metrics as consolidated metrics.

- For each phoneme category, segment-wise objective evaluation metrics mentioned were calculated and the minimum (undesirable) of each of the segment-wise objective evaluation metric value is determined. For example, if there are three vowels in a sentence, the RMSE using visual parameters is calculated for each of these segments. The maximum of the RMSE is chosen as the representative of that sentence based on a particular metric and phoneme category. This is based on the observation that, sometimes the subjective opinions can get affected by a few bad synthesis instances irrespective of a high overall

	Correlation							RMSE			Dur. Ratio	
Sen #	pc1	pc2	pc3	mfcc1	mfcc2	mfcc3	f0 Voi	PCs	MFCCs	f0 Voi.	All. Ph.	Vow.
1	0.874	0.772	0.771	0.852	0.658	0.812	0.715	19.75	27.75	83.05	0.44	0.54
2	0.948	0.851	0.885	0.866	0.503	0.772	0.853	13.10	26.71	62.93	0.23	0.24
3	0.926	0.910	0.824	0.900	0.659	0.775	0.756	13.34	25.91	85.24	0.58	0.37
4	0.924	0.885	0.883	0.858	0.728	0.630	0.786	11.83	24.51	81.37	0.38	0.55
5	0.946	0.834	0.899	0.874	0.627	0.870	0.914	14.77	24.58	50.66	0.27	0.27
6	0.845	0.644	0.826	0.794	0.707	0.768	0.838	14.83	28.65	65.85	0.42	0.44
7	0.912	0.887	0.746	0.867	0.504	0.782	0.837	13.32	26.95	74.14	0.50	0.78
8	0.882	0.305	0.849	0.910	0.658	0.872	0.843	13.33	23.49	60.24	0.32	0.35
9	0.855	0.536	0.627	0.686	0.363	0.809	0.597	14.99	30.15	111.38	0.65	1.02
10	0.831	0.480	0.762	0.863	0.640	0.805	0.833	12.50	26.45	69.49	0.27	0.29
11	0.946	0.932	0.886	0.849	0.724	0.819	0.857	11.27	25.55	59.42	0.47	0.57
12	0.926	0.846	0.799	0.929	0.625	0.860	0.907	13.61	24.42	50.47	0.42	0.54
13	0.908	0.870	0.851	0.688	0.469	0.731	0.601	11.38	29.27	129.75	0.42	0.37

Table 6.7: Objecting evaluation results for the out-of-corpus sentences. Vow. is for vowels, Ph. is for phonemes, Voi. is for voiced phonemes, mfcc is for Mel-frequency cepstral coefficients, PC is for principal component. The unit of f0 is Mel.

performance. We refer to these metrics as worst-case-based metrics.

With these objective metrics calculated, the subjective evaluation results for Q1 (AV synchrony), Q3 (acoustic speech naturalness) and Q4 (visual naturalness) were correlated. This was an attempt to investigate the influential aspects which drive the perceptual opinion about the synthesized speech. The correlation results suggest the possibility of the following relations:

- A correlation between Q1 scores (synchrony) and visible-vowels. This observation is based on Q1 scores and the consolidated correlation coefficients in visual and acoustic modality for visible-vowels.
- A correlation between Q3 scores (acoustic speech naturalness) and worst-case acoustic segments. This observation is based on the Q3 scores and worst-case-based acoustic speech correlation.
- A correlation between Q3 scores and vowel durations. This observation is based on the Q3 scores and consolidated vowel duration metrics. Vowels are known to be important for prosodic patterns.
- A correlation between Q4 scores (visual speech naturalness) and vowels and semi-vowels. This observation is based on the Q4 scores and the consolidated visual speech correlations for vowels and semi-vowels.
- A correlation between Q4 scores and voiced-invisible phonemes. This observation is based on the Q4 scores and consolidated correlation of visual speech for voiced-invisible phonemes. This is probably due to human beings being critical towards coarticulation.

This was just a preliminary experiment to investigate for informative patterns. But to draw definite conclusions, more rigorous systematic experiments are necessary. This kind of analysis for the intelligibility results is planned for the future.

6.4 Conclusion

In this chapter, we have described the various automatic and human-centered evaluation techniques that we have used to evaluate our system. The former techniques include correlation, RMSE calculated based on acoustic and visual parameters and duration related metrics. We have used them for evaluating various methodologies for improving selection during the development of the system. The latter, i.e., perceptual evaluation through human participants was done for the overall evaluation of the final system. Our focus was to synthesize the articulatory

dynamics. The overall evaluation results show that the synthesis is of reasonably good quality though there is still scope for improvement. The results show that we have achieved the objective of synthesizing the articulatory dynamics reasonably well[6].

[6]Parts of this chapter were published in (Musti et al., 2012), (Toutios et al., 2011).

Chapter 7

Conclusion

The work presented in this thesis deals with audio-visual speech synthesis. Our goal was to develop a system which synthesizes perfectly aligned audio-visual speech with a dynamics closer to natural speech. This is the first important step towards the development of a talking-head. For synthesis, we choose unit selection paradigm which is a corpus based concatenation framework. To avert the audio-visual alignment problem completely, we keep the natural association between acoustic and visual modalities intact. The first requirement to implement the idea was to have a synchronous bimodal speech corpus. This required corpus was acquired using a stereo-vision based motion capture technique developed by members in team MAGRIT. The bimodal speech corpus consisted of 3D point trajectories along with the corresponding synchronous audio. The face is represented as a sparse mesh using these 3D points describing the outer surface of the face. To begin with, two necessary steps needed to be accomplished. First, bimodal speech database need to be prepared using the recorded corpus. Second, we required a basic acoustic-visual speech synthesis system, which would implement the central idea to synthesize bimodal speech for a given text using the database. We processed the 3D marker data to reduce noise by applying a low pass filter. Subsequently we reduced the dimensionality of the visual modality by applying principal component analysis. We also extracted labial articulatory features from the data for further analysis. The visual data is stored as the PCA coefficients to be reprojected on to original space for facial animation.

The recorded bimodal speech corpus is a valuable resource for mining interesting information regarding speech articulation and interaction between the two modalities, which is important for speech synthesis. It's informative to study the data, its advantages and its limitations. As a start of our corpus processing, we started with segmentation experiments. We performed visual speech segmentation using the facial marker data. In fact, acoustic speech is the result of coordinated movement of articulators. Thus, the vocal tract has to take the necessary configuration in

advance for the generation of a particular sound. So, we investigated this relationship, and measured the time differences between the visual and acoustic segment boundaries. The results of these experiments were informative in planning the later steps. Firstly, it indicated the component of visual speech related information that was present in the facial data alone, without the internal vocal tract information. Subsequently, we performed segmentation experiments using EMA data which had the labial and tongue related information. The results of these automatic segmentation gave us an estimation of the missing perceptual information due to the lack of tongue in our facial data. The results of these experiments, without and with tongue related information are in agreement to the order of results shown in (Yehia et al., 1998). It indicated that the effect of missing tongue information in the visual speech is not very high and hence the resultant visual speech might still be intelligible. It would be interesting to explore in the future the possibility of labeling candidates in terms of suprasegmental features in the corpus based on such segmentation results.

For the database preparation for our system, we first performed speech segmentation using acoustic speech and took the boundaries to represent the segment boundaries in both acoustic and visual modality. This allows the possibility of keeping the association of acoustic and visual modality intact besides keeping the representation of segments simple and straightforward. The synthesis unit in our system is diphone and this choice is good for many reasons. First, the diphone includes the region of coarticulation between two neighboring phonemes. It thus also includes the visual and acoustic segment boundaries. This is the second advantage especially when we are dealing with two modalities. Third, the acoustic speech signal is relatively stationary in the middle of the phoneme. This is the point of concatenation when diphone is a synthesis unit which improves the probability of good concatenation without perceptual discontinuity. For the development of the initial basic framework of acoustic-visual speech synthesis, we started with an acoustic speech synthesizer SoJA (Colotte, 2009). Using the tools that were developed under the framework of SoJA, we segmented the acoustic data and built the speech database.

Synthesis results using unit selection depend on the various cost functions involved and their correlation to human perception. We built the system to select bimodal segments initially using target features which are extracted through text analysis alone. The synthesis segments were selected by minimizing a combination of cost functions, including the concatenation costs in the visual and acoustic domains. The concatenation cost in the acoustic domain was based on Kullback-Leibler divergence calculated using LPC coefficients. This choice was made by considering the available literature about discontinuity perception and objective distance measures. The concatenation cost in the visual domain was squared Mahalanobis distance calculated using

PCA coefficients. This overall framework of acoustic-visual speech synthesis provided the interesting ground to experiment with various methodologies for improving the synthesis performance further.

There were three domains where improvement was obviously possible. First, the set of target features which were purely based on the text analysis needed to be refined to take the corpus specific characteristics into account. Especially in the case of visual modality, the target features need to take into account the speaker-specific articulatory information accurately. Without this, the coarticulation of the synthesized speech might show perceptual incoherence to users. Hence, we developed visual target features to take this available information from the corpus accurately. We developed visual target costs based on the specific features which seem to be affected by the contexts rather than based on the context. We reported the objective evaluation metrics which show marginal improvement. This can be attributed to the large target features set in which the relative importance of the introduced feature is only about 1%.

Besides a good target and candidate description in terms of target features, the weighting of the complete set of target features in the order of their relative importance is necessary. This serves as the basis for the optimal corpus usage. Generally, unit selection based speech synthesis systems are developed on a specific set of target features. Little consideration is given in reviewing the relevance of those features explicitly, once they are manually chosen. The relative importance is implicitly taken into account through the weighting process. Unlike this approach, we developed an algorithm to explicitly perform redundant target feature elimination and simultaneously weighting the important target features. The evaluation of a target cost is done by comparing the ordering given by it and the ordering given by a distance metric based on actual speech comparison (bimodal). The relative weight given to each target feature depends on the information it contributes with its presence in the target cost compared to its absence. A feature is eliminated if its inclusion actually increases confusion in the ordering. The algorithm is robust and reasonably insensitive to the initial conditions. This way of feature selection is advantageous as high dimensionality reduces the probability of perfect candidate with exact match thus might introduce noise. This problem is alleviated to a large extent by feature selection. The distance measure used for comparing two speech realizations in the above algorithm includes four constituents. These four constituents roughly represent duration, pitch, local acoustic and visual features. The selected features and their relative importance are in good agreement to the phonetic and linguistic studies. For example, syllable position in rhythm group has shown to be the most important feature for the prediction of duration. These observations show the strength of the algorithm .

This weight tuning approach might benefit from the following investigation. Firstly, the dissimilarity measure used for the comparison of two speech realizations might be further refined by considering different constituent metrics. There are studies available which investigate various distance metrics for estimating the concatenation cost with respect to their correlation with human perception (Wouters and Macon, 1998; Vepa et al., 2002; Klabbers and Veldhuis, 1998). Similar studies for developing distance measures for comparing two speech segments will contribute to better speech synthesis. Secondly, the weights given to these constituent metrics can be further improved by systematic perceptual experiments with human participants. It can be argued that this process is inefficient and slow. But, a good justification to such an approach is that weight tuning problem in the huge dimensional space of target features is being mitigated by setting the weights in the dissimilarity metric which is of a much smaller dimension. Also, since the synthesized speech is targeted for humans, reinforcement from human participants is advantageous. Thirdly, the importance of the different aspects of dissimilarity metric varies among phonemes. For example, it is known that vowel durations are important for good prosody. But the relative importance of different aspects is kept same for all the phonemes. Besides target cost function this is true for various cost functions used for the final selection. It is known that different phonemes hold different level of importance for various factors. For example, concatenation in the middle of a vowel is more perceived to concatenation in a consonant (Syrdal, 2001, 2005). Hence, in the total cost function different phonemes have to be given different weights for various cost functions. Currently, this approach applied through methods like Weight Space Search (Hunt and Black, 1996), but it is not closely based on human perception. Though it requires drastic effort, this is an important area where dramatic improvement might be possible. The exploration might be based on a thorough survey of phonetic studies. Our experience of total cost tuning strongly suggests that this is a place where manual tuning is preferable to automatic weighting algorithms unlike target cost function. In the case of target cost function, the number weights to be set is high and it is practically inappropriate to perform manual tuning. But for total cost function, while separately tuning only the target and total cost weights, the dimensionality is low, practically feasible. Similar to any task with human intervention is tedious and time taking, it is recommended in terms of better perceptual results. This is especially true for a phoneme independent approach.

The relatively smaller size of the corpus constraints the performance of the weight tuning algorithm. Though this corpus is of smaller size when compared to a typical acoustic corpus, it is much bigger than contemporary visual speech corpora. We are planning to acquire a bigger corpus which might pave way towards further improvement in the synthesis results. But, there

are various difficulties in acquiring a big audio-visual corpus. The number of sentences which can be recorded in one day is limited. Since our corpus acquisition is based on painted markers on the face, it is important to record speech on a single day. This is because, the exact positioning of the markers on different days is difficult to ensure. Speaker-exhaustion also needs to be considered as it might affect speech utterance.

To assess the performance of our system we performed word-level perceptual intelligibility tests of our system through voluntary participants. We synthesized 1 or 2 syllabic words using our system and presented the audio-visual speech as stimulus. The underlying audio was degraded by the addition of noise to make participants pay attention to both the modalities. We also included some words present in the corpus during synthesis. These were included to estimate the highest intelligibility possible through our bimodal speech data. The intelligibility results of in-corpus words were less than those compared to a real video of person talking. This was anticipated as the face model doesn't include tongue and teeth yet. It can be said that these results are implicitly similar to those of automatic visual segmentation (chapter 4) with and without tongue data. Besides tongue and teeth being absent, the face is presented using a sparse mesh without any texture information. The results on out-of-corpus words indicate that we have been able to achieve our goal of synthesizing speech dynamics reasonably well. It can still be said that there is further scope for improvement. We believe that finding better metrics to evaluate the audio-visual speech synthesis is the key to drastically improve these systems. Both perceptual evaluations and automatic objective evaluation should be tied to enable simultaneous assessment of a synthesis system both automatically and quantitatively, and to ensure that such results are by and large coherent with human perception. We attempted establishing relation between perceptual and objective evaluation metrics. More systematic exploration is required in the future in this direction.

Publications

Parts of the work presented in this thesis were published in the following:

1. Utpala Musti, Caroline Lavecchia, Vincent Colotte, Slim Ouni, Brigite Wrobel-Dautcourt, Marie-Odile Berger (2012), ViSAC: Acoustic-Visual Speech Synthesis: The system and its evaluation, FAA: The ACM 3rd International Symposium on Facial Analysis and Animation, Vienna, Austria.

2. Utpala Musti, Vincent Colotte, Asterios Toutios, Slim Ouni(2011), Introducing Visual Target Cost within an Acoustic-Visual Unit-Selection Speech Synthesizer, International Conference on Auditory-Visual Speech Processing (AVSP 2011), Volterra, Italy, September 2011.

3. Utpala Musti, Asterios Toutios, Slim Ouni, Vincent Colotte, Brigitte Wrobel-Dautcourt, Marie-Odile Berger (2010), HMM-based Automatic Visual Speech Segmentation Using Facial Data, Interspeech 2010, Makuhari, Japan.

4. Asterios Toutios, Utpala Musti, Slim Ouni, Vincent Colotte: Weight Optimization for Bimodal Unit-Selection Talking Head Synthesis, Interspeech 2011, Florence, Italy, August 2011.

5. Asterios Toutios, Utpala Musti, Slim Ouni, Vincent Colotte, Brigitte Wrobel-Dautcourt, Marie-Odile Berger (2010), Setup for Acoustic-Visual Speech Synthesis by Concatenating Bimodal Units, Interspeech 2010, Makuhari, Japan.

6. Asterios Toutios, Utpala Musti, Slim Ouni, Vincent Colotte, Brigitte Wrobel-Dautcourt, Marie-Odile Berger (2010), Towards a True Acoustic-Visual Speech Synthesis, International Conference on Auditory-Visual Speech Processing (AVSP 2010), Kanagawa, Japan.

Appendix A

Stimulus for Perceptual and Subjective Evaluation

Table A.1: **Words used for intelligibility tests**

In-corpus words			
1.	chien	7.	chose
2.	cuisine	8.	fable
3.	presse	9.	gaz
4.	jeune	10.	maillot
5.	plaisir	11.	pied
6.	poche	12.	

Out-of-corpus words			
1.	anneau	21.	chasse
2.	grue	22.	nappe
3.	raison	23.	pousse
4.	bave	24.	néant
5.	riche	25.	beige
6.	laine	26.	chance
7.	niche	27.	rime
8.	beurre	28.	langue
9.	pelle	29.	case
10.	rite	30.	bien
11.	dalle	31.	latte
12.	rode	32.	mousse
13.	botte	33.	drap
14.	pince	34.	rouge
15.	bouche	35.	menthe
16.	rude	36.	brun
17.	fade	37.	mille
18.	oser	38.	gaffe
19.	molle	39.	cage
20.	gris		

Table A.2: **Sentences used in subjective evaluation of quality**

In-corpus sentences	
1.	Le Griffon leva ses deux pattes pour manifester sa surprise.
2.	Il était alors recordman du monde du quart de mile.
3.	Europe 1 revient deux fois sur le sujet.
4.	Une société qui fait de nos enfants des voyous.
5.	Il semble qu'il y ait eu un problème de connexion.
6.	La fillette regarda le banc des jurés.
7.	La fillette regarda le banc des jurés.

Out-of-corpus sentences	
1.	Annie s'ennuie loin de mes parents.
2.	Leur chienne a hurlé toute la nuit.
3.	Le bouillon fume dans les assiettes.
4.	Le caractère de cette femme est moins calme.
5.	Le tapis était élimé sur le bord.
6.	La vaisselle propre est mise sur l'évier.
7.	Je suis resté sourd à ses cris.
8.	Ma partition est sous ce pupitre.
9.	Ces légendes me rappellent les temps anciens.
10.	Vous avez du plaisir à jouer avec ceux qui ont un bon caractère.
11.	On dit que l'essor de ce village est important.
12.	La poire est un fruit à pépins.
13.	Je ne veux pas que vous le changiez pour le moment.

Bibliography

F. Alías and X. Llorà. Evolutionary weight tuning based on diphone pairs for unit selection speec synthesis. In *EUROSPEECH*, 2003. 25, 26, 31

F. Alías, X. Llorà, I. I. Sanz, J. C. Socoró, X. Sevillano, and L. Formiga. Perception-guided and phonetic clustering weight tuning based on diphone pairs for unit selection tts. In *INTERSPEECH*, 2004. 25

E. S. Andersen, A. Dunlea, and L. S. Kekelis. Blind children's language: resolving some differences. *Journal of Child Language*, 11(03):645–664, 1984. 7

G. Bailly, M. Bérar, F. Elisei, and M. Odisio. Audiovisual speech synthesis. *International Journal of Speech Technology*, 6:331–346, 2003. 12

G. Bailly, O. Govokhina, F. Elisei, and G. Breton. Lip-synching using speaker-specific articulation, shape and appearance models. *EURASIP J. Audio, Speech and Music Processing*, 2009. 16, 22, 31, 32

M. Balestri, A. Pacchiotti, S. Quazza, P. L. Salza, and S. Sandri. Choose the best to modify the least: a new generation concatenative synthesis system. In *EUROSPEECH*, 1999. 21

J. P. Barker and F. Berthommier. Evidence of correlation between acoustic and visual features of speech. In *International Congress of Phonetic Sciences*, 1999. 57

J. R. Bellegarda. A novel discontinuity metric for unit selection text-to-speech synthesis. In *ISCA Speech Synthesis Worksop*, 2004. 28

C. Benoît, M. Grice, and V. Hazan. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication*, 18(4):381–392, 1996. 33, 34, 35

J. Beskow. Rule-based visual speech synthesis. In *EUROSPEECH*, 1995. 12, 14

J. Beskow. Trainable articulatory control models for visual speech synthesis. *International Journal of Speech Technology*, 7(4):335–349, 2004. 14

A. Black and P. Taylor. Automatically clustering similar units for unit selection in speech synthesis. In *EUROSPEECH*, 1997. 23

A. W. Black. Perfect synthesis for all of the people all of the time. In *IEEE Workshop on Speech Synthesis*, 2002. 35

A. W. Black and K. Tokuda. The Blizzard Challenge - 2005: Evaluating corpus-based speech synthesis on common datasets. In *INTERSPEECH*, 2005. 34

V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *SIGGRAPH*, 1999. 13

O. Boëffard. Variable-length acoustic units inference for text-to-speech synthesis. In *INTERSPEECH*, 2001. 21

M. Brand. Voice puppetry. In *SIGGRAPH*, 1999. 15

C. Bregler, M. Covell, and M. Slaney. Video rewrite: driving visual speech with audio. In *SIGGRAPH*, 1997. 13, 15, 16, 22, 24, 29, 33

T. Chen and R. R. Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, 1998. 9

R. A. J. Clark, K. Richmond, and S. King. Festival 2 – build your own general purpose unit selection speech synthesiser. In *ISCA workshop on speech synthesis*, 2004. 21

R. A. J. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King. Statistical analysis of the Blizzard Challenge 2007 listening test results. In *The Blizzard Challenge 2007 workshop*, 2007. 28, 35

M. M. Cohen and D. W. Massaro. Modeling coarticulation in synthetic visual speech. *Models and techniques in computer animation*, 92:139–156, 1993. 12, 14, 15, 25

V. Colotte. Soja: French text-to-speech synthesis system, 2009. 44, 114

V. Colotte and R. Beaufort. Linguistic features weighting for a Text-To-Speech system without prosody model. In *INTERSPEECH*, 2005. 23, 25, 27

G. Coorman, J. Fackrell, P. Rutten, , and B. Van Coile. Segment selection in the l & h realspeak laboratory tts system. In *International Conference on Spoken Language Processing*, 2000. 23, 24, 25, 29

T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *European Conference on Computer Vision*, 1998. 13

E. Cosatto and H. Graf. Sample-based synthesis of photo-realistic talking heads. In *Computer Animation*, 1998. 15, 16

E. Cosatto and H. P. Graf. Photo-realistic talking-heads from image samples. *IEEE Transactions on Multimedia*, 2(3):152–163, 2000. 16

P. Cosi, D. Falavigna, and M. Omologo. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. In *EUROSPEECH*, 1991. 22

P. Cosi, E. Magno C., G. Perin, and C. Zmarich. Labial coarticulation modeling for realistic facial animation. In *International Conference on Multimodal Interaction*, 2002. 14

P. Cosi, A. Fusaro, and G. Tisato. LUCIA a new Italian talking-head based on a modified cohen-massaro's labial coarticulation model. In *INTERSPEECH*, 2003. 12

S. Demange and S. Ouni. Continuous episodic memory based speech recognition using articulatory dynamics. In *INTERSPEECH*, 2011. 63

B. Dodd. Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, 11(4):478–484, 1979. 8

R. E. Donovan. A new distance measure for costing spectral discontinuities in concatenative speech synthesis. In *ISCA Tutorial and Research Workshop on Speech Synthesis*, 2001. 28

E.Cosatto, G. Potamianos, and H. P. Graf. Audio-visual unit selection for the synthesis of photo-realistic talking-heads. In *IEEE International Conference on Multimedia and Expo*, 2000. 13, 15, 22, 24, 30, 33

J. D. Edge, A. Hilton, and P. Jackson. Model-based synthesis of visual speech movements from 3D video. *EURASIP J. Audio Speech Music Process.*, 2009:4:2–4:2, January 2009. 16

P. Ekman and W. Friesen. *Facial Action Coding System (FACS): A Technique for the Measurement of Facial Action.* Consulting Psychologists Press, 1978. 12

F. Elisie, M. Odisio, G. Bailly, and P. Badin. Creating and controlling video-realistic talking heads. In *Proceedings of the Workshop on Audio-Visual Speech Processing*, 2001. 13, 15

T. Ezzat and T. Poggio. Miketalk: a talking facial display based on morphing visemes. In *Computer Animation*, 1998. 13

T. Ezzat, G. Geiger, and T. Poggio. Trainable videorealistic speech animation. *ACM Trans. Graph.*, 21(3):388–398, July 2002. 15, 16

S. Fagel. Joint audio-visual units selection - the JAVUS speech synthesizer. *International Conference on Speech and Computer*, 2006. 17, 30, 33

G. Fairbanks. Test of phonemic differentiation: The rhyme test. *The Journal of Acoustical Society of America*, 30(7):596–600, 1958. 33, 34

C. G. Fisher. Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11(4):796–804, 1968. 14

M. Fraser and S. King. The Blizzard Challenge 2007. In *The Blizzard Challenge 2007 workshop*, 2007. 34

G. Geiger, T. Ezzat, T. Poggio, and A. M. Feburary. Perceptual evaluation of video-realistic speech. In *CBCL Paper 224/AI Memo 2003-003, MIT*, 2003. 16, 33

O. Govokhina, G. Bailly, G. Breton, and P. Bagshaw. A new trainable trajectory formation system for facial animation. In *ISCA Workshop on Experimental Linguistics*, 2006. 16

O. Govokhina, G. Bailly, and G. Breton. Learning optimal audiovisual phasing for a HMM-based control model for facial animation. In *ISCA Workshop on Speech Synthesis*, 2007. 22

K. W. Grant and S. Greenberg. Speech intelligibility derived from asynchrounous processing of auditory-visual information. *Workshop on Audio-Visual Speech Processing*, 2001. 8

K. W. Grant and V. Van Wassenhove. Detection of auditory (cross-spectral) and audio-visual (cross-modal) synchrony. *Speech Communication*, 44(1–4):43–53, 2004. 8

K. P. Green and P. K. Kuhl. The role of visual information in the processing of place and manner features in speech perception. *Perception and Psychophysics*, 45(1):34–42, 1989. 8

K. P. Green and P. K. Kuhl. Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, 17(1):278–288, 1991. 8

A. Hallgren and B. Lyberg. Visual speech synthesis with concatenative speech. In *Auditory-Visual Speech Processing*, 1998. 17, 22

P. Hoole and N. Nguyen. Electromagnetic articulography in coarticulation research. In W.H Hardcastle and N. Hewlett, editors, *Coarticulation: Theory, Data and Techniques*, pages 260–269. Cambridge University Press., 1999. 62

A. S. House, C. Williams, M. H. L. Hecker, and K. D. Kryter. Psychoacoustic speech tests: A modified rhyme test. *The Journal of the Acoustical Society of America*, 35(11), 1963. 33, 34

F. J. Huang, E. Cosatto, and H. P. Graf. Triphone based unit selection for concatenative visual speech synthesis. In *International Conference on Acoustics, Speech and Signal Processing*, 2002. 31

A. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *International Conference on Acoustics, Speech and Signal Processing*, 1996. 20, 21, 23, 25, 26, 27, 31, 87, 116

N. Iwahashi, N. Kaiki, and Y. Sagisaka. Concatenative speech synthesis by minimum distortion criteria. In *International Conference on Acoustics, Speech and Signal Processing*, 1992. 20, 27

K. Johnson, P. Ladefoged, and M. Lindau. Individual differences in vowel production. *J. Acoust. Soc. Am.*, 94(2 Pt 1):701–714, 1993. 74

V. Karaiskos, S. King, R. A. J. Clark, and C. Mayo. The blizzard challenge 2008. In *The Blizzard Challenge 2008 workshop*, 2008. 34

S. King and V. Karaiskos. The blizzard challenge 2009. In *The Blizzard Challenge 2009 workshop*, 2009. 34

S. King and V. Karaiskos. The blizzard challenge 2010. In *The Blizzard Challenge 2010 workshop*, 2010. 34

S. King and V. Karaiskos. The blizzard challenge 2011. In *The Blizzard Challenge 2011 workshop*, 2011. 34

E. Klabbers and R. Veldhuis. On the reduction of concatenation artefacts in diphone synthesis. In *International Conference on Speech and Language Processing*, 1998. 28, 100, 116

E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Transactions on Speech and Audio Processing*, 9(1):39–51, 2001. 28

J. Kominek and A. W. Black. A family-of-models approach to HMM-based segmentation for unit selection speech synthesis. In *INTERSPEECH*, 2004. 22

P. Ladefoged. *A Course in Phonetics*. Harcourt Brace Jovanovich, 2 edition, 1982. 73

P. Ladefoged and I. Maddieson. *The Sounds of the World's Languages*. Wiley-Blackwell, 1995. 73

L. Latacz, W. Mattheyses, , and W. Verhelst. The VUB Blizzard Challenge 2010 Entry: Towards automatic voice building. 2010. 23, 25, 26, 31

L. Latacz, W. Mattheyses, and W. Verhelst. Joint target and join cost weight training for unit selection synthesis. In *INTERSPEECH*, 2011. 83, 88

Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *SIGGRAPH*, 1995. 12

B. LeGoff, T. Guiard-Marigny, M. Cohen, and C. Benoit. Real-time analysis-synthesis and intelligibility of talking faces. In *International Conference on Speech Synthesis*, 1994. 7, 33

S. Lemmetty. *Review of Speech Synthesis Technology*. 1999. URL http://www.acoustics.hut.fi/publications/files/theses/lemmetty_mst/index.html. 33

W. Lijuan, Q. Xiaojun, H. Wei, and K. S. Frank. Photo-real lips synthesis with trajectory-guided sample selection. *Speech Synthesis Workshop*, 2010. 16

K. Liu and J. Ostermann. Optimization of an Image-Based Talking Head System. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009. 15, 16, 17, 31, 33, 49

A. Ljolje and M. D. Riley. Automatic segmentation of speech for tts. In *EUROSPEECH*, 1993. 22

A. Ljolje, J. van Santen, and J. Hirschberg. *Progress in Speech Synthesis*, chapter Automatic speech segmentation for concatenative inventory selection, pages 305–312. Springer-Verlag, 1997. 22

A. Löfqvist. Speech as audible gestures. In W.J. Hardcastle and A. Marchal, editors, *Speech Production and Speech Modelling*, pages 289–322. Dordrecht: Kluwer Academic Publishers, 1990. 73

J. Ma, R. Cole, B. Pellom, W. Ward, and B. Wise. Accurate visible speech synthesis based on concatenating variable length motion capture data. *IEEE Transactions on Visualization and Computer Graphics*, 12(2):266–276, 2006. 16, 30, 31, 32

S. Maeda. Compensatory articulation in speech: analysis of x-ray data with an articulatory model. In *EUROSPEECH*, 1989. 74

D. Massaro. Embodied agents in language learning for children with language challenges. *International Conference on Computers Helping People with Special Needs*, 2006. 8

T. Masuko, T. Kobayashi, M. Tamura, J. Masubuchi, and K. Tokuda. Text-to-visual speech synthesis based on parameter generation from HMM. In *International Conference on Acoustics, Speech and Signal Processing*, 1998. 15

W. Mattheyses, L. Latacz, and W. Verhelst. On the importance of audiovisual coherence for the perceived quality of synthesized visual speech. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009. 8, 17, 33

W. Mattheyses, L. Latacz, and W. Verhelst. Optimized photorealistic audiovisual speech synthesis using active appearance modeling. In *Auditory-Visual Speech Processing*, 2010. 25, 79

H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264:746–748, 1976. 9

Y. Meron and K. Hiros. Efficient weight training for selection based synthesis. In *EUROSPEECH*, 1999. 25, 26, 31

S. Minnis and A. Breen. Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with concatenative synthesis. *Sixth International Conference on Spoken Language Processing*, 2000. 15

B. Möbius. Corpus-based speech synthesis: methods and challenges. *Arbeitspapiere des Instituts für Maschinelle Sprachverarbeitung (Univ. Stuttgart), AIMS*, 6(4), 2000. 20

E. Moulines and F. Charpentier. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6):453–467, 1990. 44

U. Musti, A. Toutios, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, and M.-O. Berger. HMM-based automatic visual speech segmentation using facial data. In *INTERSPEECH*, 2010. 69

U. Musti, A. Toutios, V. Colotte, and S. Ouni. Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer. *Workshop on Audio-Visual Speech Processing*, 2011. 100

U. Musti, C. Lavecchia, V. Colotte, S. Ouni, B. Wrobel-Dautcourt, and M.-O. Berger. Visac : Acoustic-visual speech synthesis: The system and its evaluation. *FAA: The ACM 3rd International Symposium on Facial Analysis and Animation*, 2012. 101, 111

M. Odisio, G. Bailly, and F. Elisei. Tracking talking faces with shape and appearance models. *Speech Communication*, 44(1-4):63–82, 2004. 57

S. E. G. Öhman. Numerical model of coarticulation. *Journal of the Acoustical Society of America*, 41(2):310–320, 1967. 14

J.-L. Olives, R. Möttönen, J. Kulju, and M. Sams. Audio-visual speech synthesis for Finnish. *Auditory-visual Speech Processing Workshop*, 1999. 12

J. Ostermann. Animation of synthetic faces in MPEG-4. In *Computer Animation*, 1998. 13

S. Ouni, M. Cohen, H. Ishak, and D. Massaro. Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. *EURASIP Journal on Audio, Speech, and Music Processing*, 2007:47891, 2007. 7, 33

E. Owens and B. Blazek. Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech and Hearing Research*, 28(3):381 - 393, 1985. 24

I. S. Pandzic, J. Ostermann, and D. R. Millen. User evaluation: synthetic talking faces for interactive services. *The Visual Computer*, 15(7–8):330–340, 1999. 7

Y. Pantazis, Y. Stylianou, and E. Klabbers. Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis. In *INTERSPEECH*, 2005. 28

S. S. Park and N. S. Kim. On using multiple models for automatic speech segmentation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(8):2202–2212, 2007. 22

S. S. Park, C. K. Kim, and N. S. Kim. Discriminative weight training for unit-selection based speech synthesis. In *EUROSPEECH*, 2003. 25, 27

F. I. Parke. Computer generated animation of faces. 1972. 12

F. I. Parke. A model for human faces that allows speech synchronized animation. *Journal of Computers and Graphics*, 1(1):1–4, 1975. 12

F. I. Parke. Parametric models for facial animation. *IEEE Computer Graphics and Applications*, 2(9):61–70, 1982. 12

C. Pelachaud, N. I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(6):1–46, 1994. 14

C. Pelachaud, E. M. Caldognetto, C. Z., and P. Cosi. An approach to an Italian talking head. In *INTERSPEECH*, 2001. 12

H. R. Pfitzinger. Concatenative speech synthesis with articulatory kinematics obtained via three-dimensional electro-magnetic articulography. In *Proc. of DAGA 2005*, 2005. 64

B. T. Phong. Illumination for computer generated pictures. *Commun. ACM*, 18(6):311–317, 1975. 12

R. Prudon and C. d'Alessandro. A selection/concatenation tts synthesis system. In *ISCA Workshop on Speech Synthesis*, 2001. 23

L. J. Raphael and F. Bell-Berti. Tongue musculature and the feature of tension in english vowels. *Phonetica*, 32(1):61–73, 1975. 74

V. Robert, B. Wrobel-Dautcourt, Y. Laprie, and A. Bonneau. Inter speaker variability of labial coarticulation with the view of developing a formal coarticulation model for french. In *Auditory-Visual Speech Processing*, 2005. 43, 76

Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *International Conference on Acoustics, Speech and Signal Processing*, 1988. 20

J. L. Schwartz, F. Berthommier, and C. Savariaux. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93(2):B69–78, 2004. 7

A. Schweitzer, N. Braunschweiler, T. Klankert, B. Möbius, and B. Säuberlich. Restricted unlimited domain synthesis. In *EUROSPEECH*, 2003. 21, 28

Y. Stylianou and A. Syrdal. Perceptual and objective detection of discontinuities in concatenative speech synthesis. In *International Conference on Acoustics, Speech and Signal Processing*, 2001. 28

W. H. Sumby and I. Pollack. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26(2):212–215, 1954. 7

A. Q. Summerfield. Use of visual information for phonetic perception. *Phonetica*, 36(4–5): 314–331, 1979. 7

A. K. Syrdal. Phonetic effects on listener detection of vowel concatenation. In *INTERSPEECH*, 2001. 29, 116

A. K. Syrdal and A. Conkie. Perceptually-based data-driven join costs: comparing join types. In *INTERSPEECH*, 2005. 29, 116

K. Takeda, K. Abe, and Y. Sagisaka. On unit selection algorithms and their evaluation in non-uniform unit speech synthesis. *The ESCA Workshop on Speech Synthesis*, 1990. 20, 27

M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi. Text-to-audio-visual speech synthesis based on parameter generation from HMM. In *EUROSPEECH*, 1999. 17

P. Taylor. The target cost formulation in unit selection speech synthesis. In *INTERSPEECH*, 2006. 27

P. Taylor. *Text-to-Speech Synthesis*, chapter Unit Selection Synthesis. Cambridge University Press, 2009. 21

P. Taylor and A. W. Black. Speech synthesis by phonological structure matching. In *EUROSPEECH*, 1999. 21

T. Teinonen, R. N. Aslin, P. Alku, and G. Csibra. Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, 108(5):850–5, 2008. 7

B. Theobald. Audiovisual speech synthesis. In *International Congress on Phonetic Sciences*, 2007. 11

B.-J. Theobald, S. Fagel, G. Bailly, and F. Elisei. Lips2008: visual speech synthesis challenge. In *INTERSPEECH*, 2008. 35

D. T. Toledano, L. A. H. Gomez, and L. V. Grande. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6):617–625, 2003. 22

A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, and M.-O. Berger. Setup for acoustic-visual speech synthesis by concatenating bimodal units. In *INTERSPEECH*, 2010a. 50

A. Toutios, U. Musti, S. Ouni, V. Colotte, B. Wrobel-Dautcourt, and M.-O. Berger. Towards a true acoustic-visual speech synthesis. *International Conference on Auditory-Visual Speech Processing*, 2010b. 50

A. Toutios, U. Musti, S. Ouni, and V. Colotte. Weight optimization for bimodal unit-selection talking head synthesis. In *INTERSPEECH*, 2011. 49, 85, 103, 111

J. Vepa and S. King. Kalman-filter based join cost for unit-selection speech synthesis. In *INTERSPEECH*, 2003. 29

J. Vepa and S. King. Subjective evaluation of join cost functions used in unit selection speech synthesis. In *INTERSPEECH*, 2004. 28

J. Vepa, S. King, and P. Taylor. Objective distance measures for spectral discontinuities in concatenative speech synthesis. In *INTERSPEECH*, 2002. 28, 100, 116

K. Waters. A muscle model for animation three-dimensional facial expression. In *SIGGRAPH*, 1987. 12

K. Waters and D. Terzopoulous. A physical model of facial tissue and muscle articulation. In *SIGGRAPH Facial Animation Course Notes*, pages 130–145, 1990. 12

A. Weissenfeld, K. Liu, S. Klomp, and J. Ostermann. Personalized unit selection for an image-based facial animation system. In *Workshop on Multimedia Signal Processing*, 2005. 15, 25, 31

J. Wouters and M. W. Macon. A perceptual evaluation of distance measures for concatenative speech synthesis. In *International Conference on Spoken Language Processing*, 1998. 28, 100, 116

B. Wrobel-Dautcourt, M.-O. Berger, B. Potard, Y. Laprie, and S. Ouni. A low-cost stereovision based system for acquisition of visible articulatory data. In *Auditory-Visual Speech Processing*, 2005. 39

H. Yehia, P. Rubin, and E. Vatikiotis-Bateson. Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26(1-2):23–43, 1998. 57, 65, 114

S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Engineering Department, 2005. 56

Abstract

This work deals with audio-visual speech synthesis. In the vast literature available in this direction, many of the approaches deal with it by dividing it into two synthesis problems. One of it is acoustic speech synthesis and the other being the generation of corresponding facial animation. But, this does not guarantee a perfectly synchronous and coherent audio-visual speech.

To overcome the above drawback implicitly, we proposed a different approach of acoustic-visual speech synthesis by the selection of naturally synchronous bimodal units. The synthesis is based on the classical unit selection paradigm. The main idea behind this synthesis technique is to keep the natural association between the acoustic and visual modality intact. We describe the audio-visual corpus acquisition technique and database preparation for our system. We then present visual speech segmentation experiments that we did using the bimodal speech corpus. We present an overview of our system and detail the various aspects of bimodal unit selection that need to be optimized for good synthesis. The main focus of this work is to synthesize the speech dynamics well rather than a comprehensive talking head. We describe the visual target features that we designed. We subsequently present an algorithm for target feature weighting. This algorithm that we developed performs target feature weighting and redundant feature elimination iteratively. This is based on the comparison of target cost based ranking and a distance calculated based on the acoustic and visual speech signals of units in the corpus. Finally, we present the perceptual and subjective evaluation of the final synthesis system. The results show that we have achieved the goal of synthesizing the speech dynamics reasonably well.

Keywords: Audio-visual speech synthesis, unit selection, target cost, target feature weighting.