Introduction d'éléments paralinguistiques en synthèse vocale

Lionel Ségalen[1], *Didier Cadic*[2]

[1] Télécom Bretagne, Brest, France
[2] Orange Labs, TECH/SSTP/VMI, Lannion, France
lionel.segalen@enst-bretagne.fr, didier.cadic@orange-ftgroup.com

ABSTRACT

Corpus based text-to-speech systems currently produce very natural synthetic sentences, though limited to a neutral inexpressive speaking style. Paralinguistic elements are some of the expressive features one would most like to introduce. In this paper, we describe a new method for introducing laughter and hesitation in synthetic speech. Thanks to a small dedicated acoustic database, this method can successfully render transitions between speech and paralinguistic elements.
Keywords: speech synthesis, paralinguistic elements, laughter, hesitation, word-ending cluster

1. INTRODUCTION

L'apparition il y a une dizaine d'années des systèmes de synthèse vocale par corpus [1, 2] a permis un gain significatif de qualité en Text-to-Speech (TTS). Leur succès repose avant tout sur l'utilisation de bases de données mono-locuteurs de plusieurs heures de parole, contenant un ensemble très riche d'unités dans de nombreux contextes phonétiques et prosodiques.

Les restitutions vocales offertes sont très naturelles mais restent cantonnées à un style prosodique neutre. Pour cette raison ils ne peuvent pas totalement répondre aux besoins qu'ils ont pourtant fait naître : dialogues naturels homme-machine, lecture de livres pour enfants, utilisation pour le doublage de films et documentaires, *etc*. Toutes ces applications nécessitent une certaine maîtrise des traits essentiels d'expressivité de la parole humaine, parmi lesquels figurent les éléments paralinguistiques. [3]

Ces éléments représentent les bruits et mimiques qui accompagnent le flux linguistique d'un locuteur et influencent, volontairement ou non, les informations véhiculées : hésitations, éclaircissements de voix, rires, éternuements, expressions faciales, mouvements des mains… La plupart de ces éléments, en particulier ceux produits par l'appareil phonatoire, s'accompagnent de déformations spécifiques sur les phonèmes environnants. Ainsi, un rire survenant au milieu d'une phrase est en général annoncé par une altération du timbre et une élévation de la fréquence fondamentale, qui tendent à persister dans la suite de la phrase. La problématique est alors la suivante : comment intégrer "sans couture" un élément paralinguistique dans un flux de parole synthétique ?

Dans cet article nous proposons une technique d'introduction réaliste d'éléments paralinguistiques vocaux dans la synthèse de parole, sur la base d'un placement manuel de ces éléments (introduction par l'utilisateur de balises spécifiques dans le texte). Nous l'appliquons ici au cas des rires et des hésitations.

Dans la section suivante nous décrivons les éléments paralinguistiques traités ainsi que notre approche. Nous détaillons en troisième section un aspect essentiel de ces travaux, à savoir le choix du corpus d'enregistrement. La quatrième section présente notre protocole d'évaluation ainsi que les résultats obtenus, et la dernière section dresse quelques enseignements et perspectives de ces travaux.

2. TECHNIQUE D'INTRODUCTION D'ÉLÉMENTS PARALINGUISTIQUES

2.1. Eléments traités

Le rire

Le rire est l'un des éléments paralinguistiques les plus importants du fait que les utilisateurs de TTS l'associent souvent à un idéal d'expressivité. Nous l'avons donc inclus dans cette étude, bien qu'il constitue un véritable écueil pour la synthèse vocale. En effet, il a de grandes répercussions sur la prosodie (pitch, énergie…) et le timbre des mots voire des phrases environnantes. En outre, la définition de ses spécificités est rendue complexe par le fait qu'il ne possède pas de caractères stéréotypés mais plutôt de nombreuses variantes [4, 5].

Les hésitations

Nous avons souhaité traiter les hésitations car elles sont fréquentes à l'oral. Bien qu'elles puissent prendre des formes très différentes en fonction du locuteur et du contexte [6], nous nous sommes restreints au type de réalisation le plus courant : le *euh* plus ou moins long et plus ou moins laryngalisé qui s'intercale généralement entre 2 mots. L'observation de ce type d'élément indique qu'il peut aussi bien être précédé d'une courte pause que prolonger sans interruption le mot qui le précède. La première manière peut aisément être restituée en synthèse vocale (simple insertion au milieu d'une pause d'un *euh* enregistré isolément) mais est inappropriée lorsqu'une certaine rapidité et fluidité sont requises. Dans cette étude, nous nous sommes donc intéressés à ce deuxième cas, plus délicat à traiter car il nécessite d'assurer un

continuum entre la phrase de synthèse et l'élément paralinguistique.

2.2. Modèle d'élément paralinguistique

Nous proposons en figure 1 un découpage en plusieurs phases d'un flux de parole contenant un élément paralinguistique.

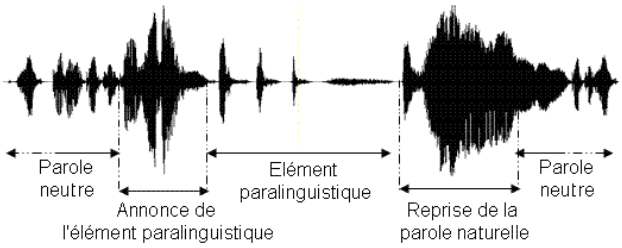

Figure 1: Segmentation d'un flux de parole contenant un élément paralinguistique (ici un rire)

Le tronçon de parole précédant l'élément paralinguistique présente des signes précurseurs, comme par exemple une augmentation du volume, de la fréquence fondamentale… Nous le qualifierons ainsi "d'annonce". Cette partie du signal varie énormément suivant le type et l'intensité de l'élément, ainsi que la nature des phonèmes. Nous définissons de la même manière la zone de "reprise de la parole", qui fait parfois l'objet d'une prolongation des caractéristiques vocales liées à l'élément paralinguistique.

La nature de ces zones transitoires est étroitement liée à l'élément lui-même. En effet, un rire très expressif et intense s'accompagnera de modifications plus importantes sur les phonèmes environnants qu'un rire contenu.

Bien entendu, ce modèle des différentes phases de la production d'un élément paralinguistique ne reflète qu'une vision simplifiée de la réalité ; certains cas de figures ne sont pas pris en compte, comme par exemple les allongements récurrents dans le cas d'une phrase hésitante.

2.3. Approche proposée

Les travaux relatés dans cet article portent exclusivement sur la production des éléments paralinguistiques et de leur phase d'annonce en synthèse par corpus. Pour les hésitations comme pour les rires, nous proposons d'enrichir notre base de données d'un certain nombre de transitions entre la parole et l'élément en question. Cette approche vise à éviter le recours à des procédés de traitement du signal pour simuler les phases transitoires, puisqu'ils tendent à dégrader le naturel du signal de parole.

Naturellement, le nombre d'enregistrements nécessaires pour cet enrichissement augmente très rapidement avec la longueur du contexte que l'on souhaite voir influencé et avec le niveau de couverture que l'on souhaite atteindre. L'annonce d'un rire peut ainsi porter uniquement sur les derniers phonèmes (explosion assez subite) ou bien s'étendre sur plusieurs mots (éclat de rire progressif). Dans la section suivante nous expliquons la stratégie adoptée pour concilier le naturel des restitutions et les contraintes de la synthèse par corpus.

Le corps de l'élément paralinguistique étant souvent difficilement détachable de son annonce, nous avons décidé de préserver cette continuité en enregistrant les deux phases d'une traite.

La cohérence avec la phase de reprise n'est en revanche pas assurée dans notre système ; nous nous sommes pour l'instant contentés de la simple insertion d'une pause entre l'élément paralinguistique et la reprise de la vocalisation.

3. CHOIX DU CORPUS ET ENREGISTREMENT

3.1. Généralités

Le choix du corpus d'enregistrement pour la couverture des éléments paralinguistiques est l'un des points essentiels de notre approche. En effet, il doit d'une part être suffisamment riche pour apporter dans la plupart des cas un continuum satisfaisant entre la phrase de synthèse et l'élément paralinguistique qui suit, et d'autre part rester le plus court possible pour des raisons de coût évidentes. Nous faisons l'hypothèse que les éléments paralinguistiques choisis ne peuvent pas survenir à l'intérieur d'un mot. Ainsi leur annonce porte nécessairement sur la fin du mot précédent, dont il s'agit donc d'optimiser la couverture. Nous avons par conséquent procédé à une étude statistique des fins de mots dans la langue française.

Pour cela nous avons introduit la notion de cluster de fin de mot qui désigne, pour un mot donné, l'ensemble des phonèmes incluant et suivant la dernière consonne du mot. Ce choix est justifié par le fait qu'en synthèse par concaténation de diphones, les consonnes, par leurs faibles énergies et cohérence temporelle, ainsi que par l'acuité réduite de leurs formants, supportent bien mieux les concaténations que les voyelles : les artefacts sont moins audibles. La dernière consonne avant l'élément paralinguistique est donc un lieu de segmentation privilégié pour le démarrage de l'annonce de l'élément. Toutefois les liquides [l] et [ʀ] (notation API), qui sont fortement influencées par leur contexte surtout lorsqu'il est non-voisé (risque de dévoisement de la liquide), n'ont pas été considérées comme segmentables lorsqu'elles étaient précédées d'un phonème sourd ([p]-[t]-[k]-[f]-[s]-[ʃ]). Le tableau 1 présente quelques exemples de clusters de fin de mot.

mot	cluster de fin
cinéma [sinema]	[ma]
remercier [ʀəmɛʀsje]	[sje]
déplacent [deplas]	[s]
peuple [pœplə]	[plə]
plâtrier [plɑtʀije]	[tʀije]

Tableau 1 : Exemples de clusters de fin de mot

3.2. Statistiques des clusters de fins de mots

Nous avons calculé les statistiques de ces clusters de fin de mot sur un corpus textuel de 130 000 groupes de souffle composé de sous-titres de films, de pièces de théâtre contemporaines, de recettes de cuisine, et d'articles du Monde. 453 548 clusters de fins de mots ont ainsi été relevés, dont seulement 3340 distincts. La figure 2 présente la fonction de répartition de ces clusters classés par ordre décroissant de fréquence d'apparition.

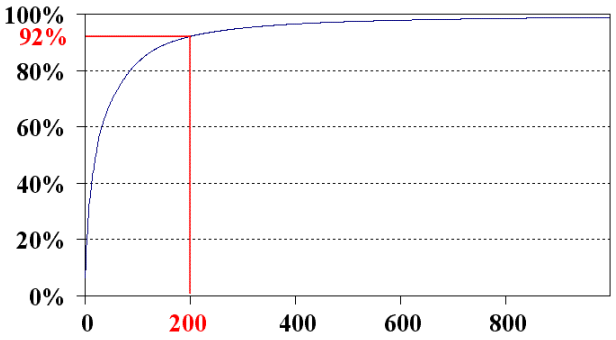

Figure 2: Proportion de fins de mots couvertes dans le corpus en fonction du nombre de clusters retenus

L'allure logarithmique de cette courbe est fréquemment observée en linguistique statistique [7] du fait de la loi de Zipf-Mandelbrot qui régit la plupart des phénomènes de ce type. Les convergences en 0 et $+\infty$ sont des effets de bords dus à la taille finie du corpus. Elle suggère ici que l'on peut obtenir une couverture de 92% des occurrences de fins de mots avec seulement 200 clusters distincts, ce qui est très satisfaisant pour notre application.

3.3. Phase d'enregistrement

Pour le rire comme pour l'hésitation, nous avons enregistré chacun des 200 clusters de fin de mot les plus fréquents, en les faisant précéder d'un contexte neutre de type logatome, et en les faisant suivre de l'élément paralinguistique en question. Un ingénieur de l'équipe a prêté sa voix pour ces enregistrements. Nous lui avons entre autres demandé de conserver une voix neutre jusqu'à la consonne initiale du cluster, ceci afin que ce dernier puisse être concaténé sans difficulté après n'importe quelle phrase de synthèse. La phase d'annonce est donc concentrée sur le cluster.

L'enregistrement des 200 prompts de rire a nécessité un peu moins de 30 minutes, de même pour les hésitations. Notons que notre locuteur amateur a éprouvé quelques difficultés à simuler les rires, ce qui souligne l'importance de recourir à un acteur professionnel pour ce type d'enregistrements.

Les 400 prompts ainsi obtenus ont été ajoutés à la base de synthèse par corpus dont nous disposions déjà pour ce locuteur (25 min de parole utile) ; nous expliquerons au paragraphe 4.2 comment nous avons compensé la petite taille de cette base. Les rires, hésitations, ainsi que leurs annonces ont été segmentés manuellement et annotés de manière spécifique dans la base. Le moteur de sélection de diphones a été modifié afin de prendre en compte ces nouveaux marqueurs. Nous ne détaillerons pas ces aspects techniques. Notons toutefois que, lors de la synthèse, l'élément paralinguistique et son annonce sont, pour les 92% de clusters couverts, introduits par une concaténation sur la dernière consonne du mot précédent. Les 8% restants peuvent tous être obtenus par concaténation sur un phonème voyelle ou semi-voyelle.

4. EVALUATION

4.1. Principe de l'évaluation

Nous nous sommes intéressés à l'évaluation subjective de notre technique de restitution d'éléments paralinguistiques, et avons opté pour des tests de type MOS (Mean Opinion Score, recommandation UIT-T P.85).

Pour cette évaluation, chaque phrase restituée par notre système a été mise en regard de deux solutions limites : d'une part une version entièrement naturelle de la même phrase incluant l'élément paralinguistique (actée au mieux par le locuteur), d'autre part une version basique obtenue par simple insertion de l'élément dans le flux de parole (au milieu d'un silence).

4.2. Préparation des phrases de test

Pour les hésitations comme pour les rires, nous avons retenu 10 phrases tests. Dans chacune d'elles, l'élément paralinguistique est entouré de deux groupes de mots, l'ensemble formant un tout sémantiquement cohérent.

Pour les versions synthétiques de ces phrases (basique / avec annonce), nous avons souhaité focaliser l'attention des auditeurs sur l'élément paralinguistique et non sur la synthèse des phrases environnantes. Ainsi, pour éviter l'apparition d'artefacts qui parasiteraient ces zones annexes, nous avons simulé une synthèse de très haute qualité en ajoutant simplement une version neutre de ces 20 phrases (sans élément paralinguistique) à la base de données de notre locuteur. Ainsi le système n'effectue que 2 concaténations, de part et d'autre de l'élément.

De plus, l'appréciation des rires étant extrêmement subjective et dans un souci d'équité, nous avons essayé de réutiliser pour la version basique les rires obtenus avec notre système, lorsqu'ils pouvaient facilement être séparés de l'annonce (souvent possible grâce aux nombreux silences qui émaillent les rires). Lorsque cela était malaisé ainsi que pour les hésitations, nous avons inséré dans la version basique des éléments paralinguistiques enregistrés isolément et qui nous semblaient les plus appropriés.

4.3. Déroulement du test

Les 60 prompts sonores (10 phrases avec 3 méthodes pour 2 types d'éléments) ont été écoutés et notés dans un ordre aléatoire par 10 auditeurs naïfs d'origine française.

Afin d'apprécier l'impact des différentes phases sur la perception, nous avons retenu trois critères de notation :

- le naturel de la phase d'annonce (transition entre la parole et l'élément paralinguistique)
- le naturel de l'élément paralinguistique lui-même
- le naturel de la reprise de vocalisation (transition entre l'élément et la parole suivante)

Les auditeurs ont noté chaque prompt suivant chacun des 3 critères, sur une échelle de 1 à 5 correspondant respectivement à : *mauvais*, *médiocre*, *passable*, *bon*, *excellent*.

4.4. Résultats

Les figures 3 et 4 ci-dessous rapportent les notes MOS obtenues respectivement pour les hésitations et pour les rires, réparties par méthode et par critère de notation. Nous y avons reporté les intervalles de confiance à 95%.

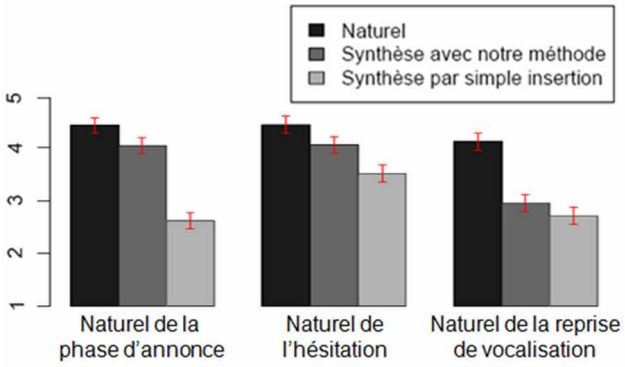

Figure 3: MOS obtenus pour les hésitations

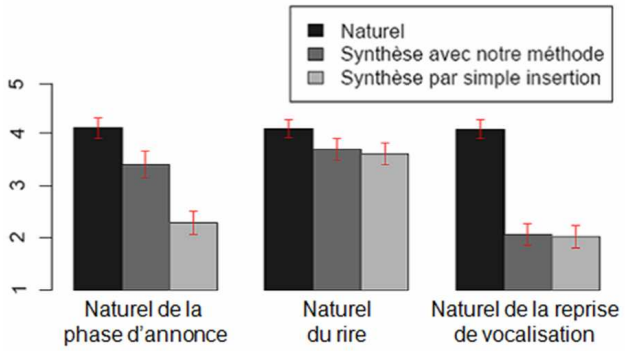

Figure 4: MOS obtenus pour les rires

Concernant la phase d'annonce, ces résultats indiquent un apport très net de notre système par rapport au procédé basique d'insertion de l'élément paralinguistique au milieu d'un silence. Pour les hésitations, nous obtenons même une note très proche du naturel. Pour le rire, notre phase d'annonce reste vraisemblablement un peu trop courte pour paraître totalement naturelle. Notons que dans 25% des cas elle ne porte que sur la consonne finale.

L'échelonnage similaire des notes pour le "naturel des hésitations" traduit la difficulté des auditeurs à décorréler les critères de notation entre eux : l'hésitation et son annonce tendent à former un tout indissociable tandis que le rire, plus hâché, s'accorde mieux avec le modèle en 3 phases.

Enfin, pour les deux types de synthèse, la reprise de vocalisation est très mal notée (rires et hésitations). Ceci est justifié par le fait que nous n'assurons absolument aucune continuité entre l'élément paralinguistique et la parole suivante, comme dans le procédé basique. Suite aux remarques des différents sujets, il semblerait que l'insertion d'une respiration juste avant la reprise de la vocalisation puisse à elle seule améliorer sensiblement leur appréciation.

5. CONCLUSION

Nous avons proposé une nouvelle méthode d'introduction d'éléments paralinguistiques vocaux en TTS, basée sur l'ajout à la base locuteur de transitions entre des fins de mots et les différents types d'éléments. Nous sommes parvenus à assurer une bonne couverture des transitions possibles avec un corpus d'enregistrement assez limité (session de 30 min).

Nous avons expérimenté cette technique sur les rires et hésitations, et l'avons comparée subjectivement à des prompts naturels et des prompts synthétiques basiques, obtenus par simple insertion d'un élément paralinguistique enregistré isolément. Les résultats montrent un réel apport de notre méthode, par rapport au procédé basique, pour la transition de la parole vers l'élément.

Ces résultats suggèrent par ailleurs d'étudier la reprise de vocalisation, soit en suivant une approche analogue, soit en envisageant plus simplement l'insertion d'une respiration. Nous menons à l'heure actuelle des travaux en ce sens. Une application à d'autres langues est également envisagée. Enfin, la réalisation de certains éléments paralinguistiques nécessitant de réels talents d'acteurs, il semble important de recourir à un locuteur professionnel pour la suite de ces expériences.

BIBLIOGRAPHIE

[1] Sagisaka Y., *Speech synthesis by rules using an optimal selection of non-uniform synthesis units*, ICASSP'88, pp. 679-682

[2] Hunt A., Black A., *Unit selection in a concatenative speech synthesis system using a large speech database*, ICASSP'96, pp. 373-376

[3] Campbell N., *Conversational speech synthesis and the need for some laughter*, TASLP 2006

[4] Emond C., *Une analyse prosodique de la parole souriante : étude préliminaire*, JEP'06, pp 147-150

[5] Trouvain J., Schröder M, *How (not) to add laughter to synthetic speech*, ADS 2004, pp. 229-232

[6] Boula de Mareuil P. et al., *A quantitative study of disfluencies in French broadcast interviews*, DiSS'05, pp. 1-6, 2005

[7] Van Santen J., *Combinatorial issues in text-to-speech synthesis*, EUROSPEECH'97, pp. 2511-2514