Détection automatique de sons bien réalisés

Yves Laprie, Safaa Jarifi, Anne Bonneau et Dominique Fohr

LORIA CNRS
Vandœuvre-lès-Nancy, France
Tél. : +33 (0)3 83 59 20 36 - Fax : +33 (0)3 83 27 83 19
Mél :laprie,bonneau,fohr@loria.fr

ABSTRACT

Given a phonetic context, sounds can be uttered with more or less salient acoustic cues depending on the speech style and prosody. In a previous work we studied strong acoustic cues of unvoiced stops that enable a very reliable identification of stops. In this paper we use this background idea again with a view of exploiting well realized sounds to enhance speech intelligibility within the framework of language learning. We thus designed an elitist learning of HMM that make very reliable phone models emerge. The learning is iterated by feeding phones identified correctly at the previous iteration into the learning algorithm. In this way models specialize to represent well realized sounds. Experiments were carried out on the BREF 80 corpus by constructing well realized phone models for unvoiced stops. They show that these contextual models triggered off in 60% of stops occurrences with an extremely low confusion rate.

1. INTRODUCTION

En fonction de contraintes articulatoires et auditives, la prononciation d'un son, quel que soit le contexte phonétique dans lequel il est émis, varie d'une articulation très soignée à une articulation très relâchée [6]. Sur le plan acoustique et perceptif, cela signifie qu'un même son dans un même contexte phonétique possède des indices plus ou moins bien marqués et un niveau d'intelligibilité très variable. Partant de l'idée qu'un trait bien marqué acoustiquement doit pouvoir être reconnu de manière très fiable en reconnaissance automatique de la parole, nous avons défini un jeu d'indices, appelés « indices forts » spécialement conçus pour l'identification de ce type de trait. Ces indices ont été définis à partir de connaissances acoustico-phonétiques et testés grâce à un système de reconnaissance à base de règles semi-automatique. Puisque le but des indices forts est de rechercher des représentants bien réalisés acoustiquement et de ne pas commettre d'erreur, ils ne sont pas déclenchés systématiquement et un paramètre important pour juger de leur efficacité est leur taux de déclenchement. Nous avons obtenu des taux de déclenchement d'environ 30% sur un corpus d'occlusives sourdes du français [1, 2].

La notion d'indices forts ne se confond pas avec celle d'indices robustes ou de landmarks, qui sont très performants mais ne recherchent pas particulièrement de « beaux représentants » d'un son ou d'un trait et qui se déclenchent systématiquement, quel que soit le son à identifier. La détection très fiable de sons bien réalisés peut avoir deux types d'applications : fournir des informations fiables en RAP d'une part, améliorer l'intelligibilité de la parole par le renforcement des sons bien réalisés, d'autre part.

Pour une application réelle, entièrement automatique, nous sommes confrontés au problème suivant. Les systèmes à base de connaissance peuvent fournir des connaissances sur la manière dont est réalisé un son ou un trait donné mais ne peuvent fonctionner de manière entièrement automatique, notamment à cause de problèmes de segmentation, d'extraction des indices acoustiques de base (formants, caractéristiques spectrales des bruits d'explosion ou de friction…) et du nombre trop importants de sources de variation à prendre en compte lors des ajustements de seuils. A l'inverse, les systèmes de reconnaissance stochastiques sont très performants et entièrement automatiques mais ne peuvent pas fournir d'information sur la réalisation d'un son donné. Nous avons mis donc au point une méthode visant à détecter les sons bien réalisés acoustiquement de manière entièrement automatique. Cette méthode et les résultats obtenus sont présentés dans les sections 2 et 3 de ce papier.

2. MÉTHODE

Nous avons choisi un système de reconnaissance stochastique, développée par D. Fohr, et fondé sur les modèles de Markov. Comme nous l'avons dit en introduction, ce type de système ne fournit aucun retour sur la manière dont s'est effectuée l'identification d'un son. Nous ne savons pas quelle probabilité d'erreur est liée à l'identification proposée, et, à fortiori, si le son est correctement réalisé. Nous avons donc mis au point une stratégie de détection destinée à forcer le système de reconnaissance à détecter les sons bien réalisés.

Voici comment nous avons procédé. Le système effectue une boucle « apprentissage-reconnaissance » sur le corpus d'apprentissage, de telle sorte que, à la fin de cette boucle, nous possèdions pour chaque son des modèles construits à partir d'exemplaires systématiquement bien identifiés, et des modèles construits à partir d'exemplaires qui ont été au moins une fois mal identifiés.

Plus précisément, à chaque itération de la boucle, les sons qui ont été bien identifiés à l'étape précédente (tous les sons à la première étape) et qui sont correctement identifiés lors de l'itération en cours sont placés dans la classe des « bons exemples » du son qu'ils représentent. Les sons qui ont été mal identifiés au moins une fois sont placés dans la classe des « mauvais exemplaires ». À chaque nouvelle itération donc, deux sortes de modèles sont créés pour chaque son : de « bons modèles » qui sont construits

à partir de sons systématiquement bien identifiés, et de « moins bon modèles » qui sont construits à partir des autres sons de cette classe. Plus nous effectuons d'itérations, plus le taux de reconnaissance obtenu par les modèles associés aux sons bien identifiés augmente, mais plus le nombre de bons exemplaires à partir desquels sont construits ces modèles diminue. En pratique, nos expériences ont montré que seules trois voire quatre itérations suffisaient (cf. section 3). Au-delà, le taux d'identification n'augmente pas beaucoup, mais le nombre de bons exemplaires baisse.

Une approche comparable a été utilisée par Schwenk [7] pour orienter l'apprentissage de perceptrons multicouches dans le cadre de la reconnaissance de la parole. Plus récemment Greenberg [3] a adopté une approche similaire dans le cadre de l'étiquetage articulatoire d'une base de données à l'aide d'une approche connexionniste. Ayant remarqué qu'une partie des trames spectrales étaient relativement mal reconnues à l'issue de l'apprentissage, il a supprimé ces trames d'une seconde phase d'apprentissage de manière à obtenir des perceptrons multicouches plus discriminants. Dans notre cas nous souhaitions nous rapprocher des indices forts que nous avions définis en ne retenant que les bonnes réalisations acoustiques. Nous avons donc poussé la phase de sélection plus loin en itérant l'étape de sélection des bons exemples. Pratiquement la sélection des bons exemples va sans doute plus loin que la simple recherche des sons bien marqués acoustiquement. En effet, la taille des corpus actuels utilisés pour l'apprentissage des modèles acoustiques est telle qu'il est impossible d'étiqueter phonétiquement tout le corpus à la main. Par conséquent, l'étiquetage est obtenu par alignement automatique entre la phrase enregistrée et la transcription phonétique attendue. Cette méthode permet l'exploitation de très vastes corpus de parole mais ne garantit pas la pertinence des exemples sur lesquels repose l'apprentissage des modèles acoustiques. L'étape de sélection des bons exemples conduit sans doute à l'élimination d'un certain nombre d'erreurs d'étiquetage phonétique.

3. MISE EN ŒUVRE

Nous avons utilisé 5327 phrases (corpus A) du corpus BREF 80 [5] pour l'apprentissage élitiste, un autre ensemble de phrases de 361 phrases (corpus B) pour les tests et pour construire une seconde version des modèles de sons bien reconnus.

Voici les étapes de l'apprentissage élitiste :

Étape 1 On effectue un premier apprentissage classique à partir de la transcription phonétique disponible ce qui conduit aux modèles acoustiques traditionnels.

Étape 2 On effectue la reconnaissance des phrases du corpus d'apprentissage à partir de ces modèles.

Étape 3 On détermine par alignement forcé avec la transcription phonétique de la phrase les sons bien reconnus. Les étiquettes des sons mal reconnus sont modifiées.

Étape 4 On effectue un nouvel apprentissage à partir des sons bien reconnus.

On répète l'apprentissage en revenant à l'étape 2 tant que le taux d'identification des bons modèles n'est pas suffisant et que le nombre d'exemples d'apprentissage le permet.

TAB. 1: Organisation des modèles suivant le contexte phonétique

Occlusive	Voyelle suivante	Modèle contextuel
/p t k/	/a / centrale	pA tA kA
	/u o ɔ/ arrière	pU tU kU
	/i e ɛ/ avant non arrondie	pI tI kI
	/y ø œ ə/ arrondie	pY tY kY
	/ã ɛ̃ õ/ nasale	pN tN kN

Nous avons construit des modèles contextuels d'occlusives en fonction de la voyelle qui suit (cf. Tab. 1) de manière à obtenir des modèles bien marqués et suffisamment caractéristiques acoustiquement. Il faut noter que le choix des classes construites découle des caractéristiques de l'étiquetage phonétique du corpus de parole. Cet étiquetage est volontairement peu précis et cela explique donc pourquoi /e/ et /ɛ/, /o/ et /ɔ/, /ø/, /ə/ et /œ/ ont été regroupés.

La Fig. 1 montre l'évolution du taux d'identification des sons considérés comme bon exemples pour quatre itérations de l'apprentissage. Comme cela est attendu le taux d'identification augmente alors que le nombre d'exemples conservés pour l'apprentissage diminue. Cependant il reste encore plus de la moitié des exemples d'apprentissage à la quatrième itération et le nombre des bons exemples tend à se stabiliser. Cela confirme qu'il existe bien une catégorie de sons suffisamment bien prononcés et/ou reconnus.

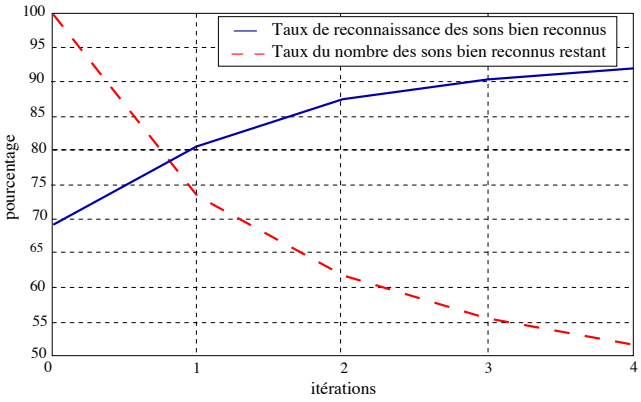


FIG. 1: Taux de reconnaissance sur le corpus d'apprentissage des sons reconnus avec les bons modèles contextuels et pourcentage restant de leurs items

Nous avons évalué les modèles de sons bien reconnus de deux façons. La première consiste à mesurer le taux de déclenchement des modèles de sons bien reconnus pour le corpus B. La Fig. 2 montre que le taux de déclenchement est en moyenne de l'ordre de 60% (en variant entre 35 et 85%) ce qui est assez élevé.

Insistons sur le fait que les chiffres de ce tableau sont les taux de déclenchement des modèles sur l'ensemble des occurrences des sons dans le corpus. Par ailleurs, il apparaît que les modèles de sons bien reconnus ne se déclenchent jamais sur d'autres classes de sons que celles pour lesquelles ils ont été conçus à l'exception de /t/ en contexte d'une voyelle antérieure déclenche très rarement sur /j/ (2,79% des cas). Cette erreur s'explique par une certaine similarité acoustique des deux sons. La seconde erreur est

la confusion du bon modèle /kA/ (/k/ dans le contexte d'un voyelle centrale) avec /tU/ (t dans le contexte d'un voyelle d'arrière). Là encore la similarité acoustique des deux sons explique cette confusion. Hormis ce cas, les confusions ont toujours lieu entre les modèles contextuels d'une même occlusive ce qui limite l'ampleur de l'erreur.

Comme pour les indices forts, le taux de déclenchement est un paramètre essentiel pour juger les « bons modèles » et nous avons obtenu un taux de déclenchement très satisfaisant pour les consonnes occlusives. Cela signifie que, pour une phrase donnée, nous pouvons dans un grand nombre de cas détecter des sons qui sont bien réalisés. C'est un résultat intéressant dans la perspective d'une application à l'apprentissage des langues.

La seconde évaluation porte sur la pertinence des modèles de sons bien reconnus. Pour cela nous avons effectué le même apprentissage élitiste sur les sons du corpus de test (corpus B) pour marquer les sons systématiquement bien identifiés de ce corpus. Ensuite nous avons mesuré le taux de déclenchement des bons modèles (obtenus sur le corpus A) sur les sons bien identifiés du corpus de test. La Fig. 3 montre que le taux de déclenchement est sensiblement plus important puisqu'il atteint 79% et que les confusions sont aussi moins nombreuses.

Ces deux expériences montrent que les modèles de sons bien reconnus correspondent bien à une classe de sons plus caractéristiques acoustiquement. Cependant rien ne garantit que ces modèles correspondent aux indices forts que nous avions définis. En effet, l'approche explicite que nous avions adoptée pour construire les indices forts conduit à un taux de déclenchement nettement plus faible, de l'ordre de 40%, et capture sans doute difficilement des cooccurrences d'indices acoustiques trouvées implicitement par l'apprentissage stochastique. Par ailleurs, nous ne disposons pas encore d'une base de sons étiquetés en indices fort suffisamment importante pour obtenir une évaluation statistiquement significative. Les tests que nous effectués montrent seulement que les modèles de sons bien reconnus ne commettent aucune erreur sur les sons repérés par les indices forts.

4. PERSPECTIVES

Une continuation logique de ce travail consiste bien entendu à tester notre méthode sur d'autres classes de sons, et nous pensons que la classe des fricatives doit également bien se prêter à la détection de sons bien réalisés.

Concernant la méthode elle-même, nous avons fait l'hypothèse que les sons systématiquement bien identifiés correspondaient à des sons bien réalisés sur le plan acoustique. Afin de vérifier cette hypothèse, nous poursuivrons donc ce travail en testant si les candidats reconnus comme des « bons exemplaires » sont effectivement des sons qui possèdent des indices acoustiques bien marqués. Nous envisageons d'ailleurs de forcer l'apprentissage des indices forts en forçant l'apprentissage à prendre en compte les sons d'un corpus pour lequel nous avons repéré les indices forts de manière semi-automatique.

Sur le plan perceptif, il nous semble particulièrement intéressant de tester si le renforcement des sons bien réalisés améliore l'intelligibilité de la parole, en particulier pour l'apprentissage des langues, ou pour les personnes souffrant de déficiences auditives. Nous avons déjà effectué des expériences de perception prouvant que le renforcement de grandes classes améliore la perception d'une langue étrangère [4]. Nous comptons poursuivre ce travail avec les sons bien réalisés.

RÉFÉRENCES

[1] A. Bonneau, S. Coste, L. Djezzar, and Y. Laprie. Two Level Acoustic Cues for Consistent Stop Identification. In *Proceedings International Conference on Spoken Language Processing*, volume 1, pages 511–514, Banff (Alberta, Canada), October 1992.

[2] A. Bonneau, S. Coste-Marquis, and Y. Laprie. Strong cues for identifying well-realized phonetic features. In *Proceedings of The XIIIth International Congress of Phonetic Sciences*, volume 4, pages 144–147, Stockholm, Sweden, 1995.

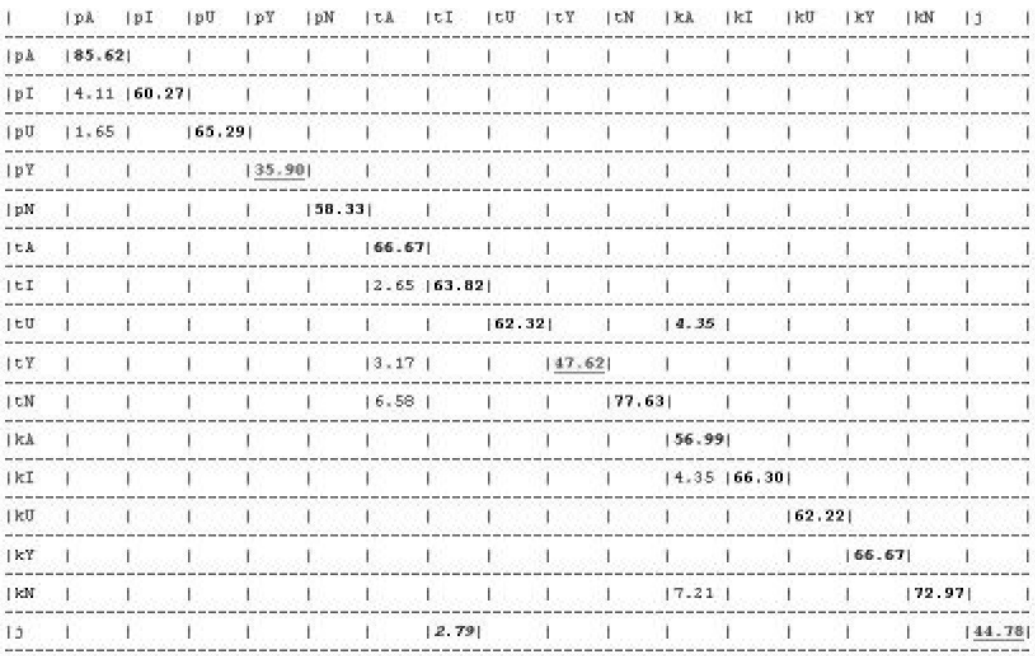
[3] S. Chang, S. Greenberg, and M. Wester. An Elitist Approach to Articulatory-Acoustic Feature Classification. In *Eurospeech, Aalborg, Danemark*, volume 4, pages 2725–2728, September 2001.

[4] Vincent Colotte, Yves Laprie, and Anne Bonneau. Signal transformation strategies to improve speech intelligibility for second language acquisition. In *17th International Congress on Acoustics, Rome, Italy*, September 2001.

[5] L.F. Lamel, J.-L. Gauvain, and M. Eskénazi. BREF, a Large Vocabulary Spoken Corpus for French. In *Proceedings of European Conference on Speech Technology*, pages 505–508, Genova, Italy, September, 1991.

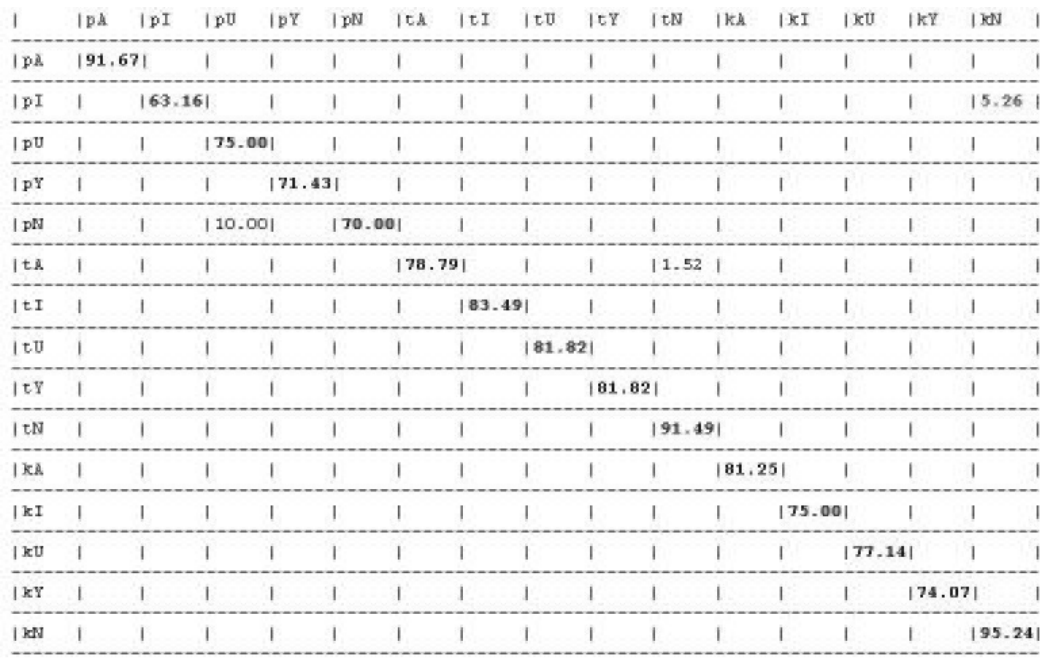
[6] B. Lindblom. Explaining phonetic variation : Sketch of the H&H theory. In W.J. Hardcastle and A. Marchal, editors, *Speech production and speech modelling*, pages 403–439. Kluwer Academic Publisher, New York, 1990.

[7] H. Schwenk. Using boosting to improve a hybrid HMM/neural network speech recognizer. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing - ICASSP'1999, Phoenix*, pages 1009–1012, Arizona, June 1999.



	pA	pI	pU	pY	pN	tA	tI	tU	tY	tN	kA	kI	kU	kY	kN	j
pA	**85.62**															
pI	4.11	**60.27**														
pU	1.65		**65.29**													
pY				35.90												
pN					**58.33**											
tA						**66.67**										
tI						2.65	**63.82**									
tU								**62.32**			**4.35**					
tY						3.17			47.62							
tN						6.58				**77.63**						
kA											**56.99**					
kI											4.35	**66.30**				
kU													**62.22**			
kY														**66.67**		
kN											7.21				**72.97**	
j							**2.79**									44.78

FIG. 2: Matrice de confusion de la reconnaissance sur le corpus B. Les chiffres de chaque colonne représentent le taux de déclenchement du bon modèle, représenté par cette colonne, sur le son représenté par la ligne. Par exemple 5.48 signifie que le bon modèle pA se déclenche sur 4.11% des sons pI



	pA	pI	pU	pY	pN	tA	tI	tU	tY	tN	kA	kI	kU	kY	kN
pA	**91.67**														
pI		**63.16**													5.26
pU			**75.00**												
pY				**71.43**											
pN			10.00		**70.00**										
tA						**78.79**				1.52					
tI							**83.49**								
tU								**81.82**							
tY									**81.82**						
tN										**91.49**					
kA											**81.25**				
kI												**75.00**			
kU													**77.14**		
kY														**74.07**	
kN															**95.24**

FIG. 3: Matrice de confusion de la reconnaissance sur le corpus B. Les chiffres de chaque colonne représentent le taux de déclenchement du bon modèle, représenté par cette colonne, sur les sons marqués comme bons exemples