Modélisation explicite de la coarticulation pour le décodage acoustico-phonétique : les triplets phonétiques

Y. Laprie, F. Lonchamp

CRIN / INRIA-Lorraine, Nancy, France
Institut de Phonétique de Nancy, Nancy, France

Abstract
This paper presents a knowledge-based approach of acoustic-phonetic decoding of continuous speech. Knowledge is stored in the form of triplets which represent contextual phone prototypes in terms of acoustic events and acoustic correlates. We describe how vowel centered triplets and plosive centered triplets are matched to the reference triplets. We then show how relaxation techniques may be used to increase the consistency of the global solution. Lastly, we indicate the aspects of the triplet approach which requires further investigation.

1 Introduction

Depuis plusieurs années l'équipe de parole du CRIN consacre une part importante de ses efforts à l'approche à bases de conaissances du décodage acoustico-phonétique. Cela a permis de développer Aphodex [Fohr 89] qui est un système expert en lecture de spectrogrammes. Cette expérience a mis en évidence qu'il est primordial de prendre en compte le contexte pour représenter la conaissance comme pour l'utiliser afin de reconnaître un son inconnu.

Cela nous a conduit à proposer un nouveau grain de conaissance pour l'approche experte en DAP : *le triplet phonétique*. Il s'agit d'un prototype de son en contexte (par exemple le son /t/ en contexte /a/ à gauche et /i/ à droite, noté $/_at_i/$) structuré en deux niveaux de description, une description acoustique en termes d'événements acoustiques de base (par exemple les formants), une composante experte faisant appel aux combinaisons d'événements acoustiques (que nous appelons indices acoustiques) reconnues significatives par l'expert (par exemple la position relative de deux formants dans un contexte phonétique donné). Pour justifier l'utilisation des triplets en DAP il faut montrer les avantages qu'ils apportent pour les points suivants :

- la précision et la pertinence de l'information acoustico-phonétique qu'ils permettent de stocker,
- la facilité de mise en œuvre de cette conaissance et notamment dans le cadre des techniques de l'intelligence artificielle qui peuvent permettre de contrôler le processus de décodage.

2 Pertinence de la représentation de la conaissance acoustico-phonétique

2.1 Représentation d'un triplet

La description acoustique d'un triplet fait appel aux événements acoustiques suivants qui peuvent être accrochés à chacune des frontières du son central d'un triplet :

- les trajectoires formantiques (valeur et pente à la frontière complétée par la valeur au centre du triplet),
- la barre d'explosion décrite par les principales concentrations d'énergie, la durée, l'intensité, les indices de compacité et de diffusion,
- bruit de friction défini par la limite inférieure de bruit et son intensité.

La composante experte regroupe l'ensemble des indices acoustiques dont la pertinence pour la reconaissance d'un triplet a été reconnue par l'expert. Ces indices sont des combinaisons d'événements acoustiques, par exemple la position relative de deux formants ou encore d'un formant par rapport à une concentration d'énergie de la barre d'explosion.

2.2 Outils destinés à l'étude de l'approche par triplet

Nous avons d'abord développé un éditeur de triplets (intégré à Snorri) qui permet à l'expert de construire les triplets directement sur le spectrogramme. Cet éditeur permet de manipuler aussi bien les triplets de référence que les instances de triplets (de la phrase à décoder) puisque leur représentation est identique. Par ailleurs, nous avons intégré à cet éditeur le module de segmentation automatique (développé dans le cadre d'Aphodex), le module de suivi automatique des formants [Laprie 90] et celui destiné à la localisation des barres d'explosion. L'utilisateur de l'éditeur peut donc s'aider de ces modules pour construire les triplets de référence qui constituent la base de conaissances.

2.3 Identification de triplets vocaliques et occlusifs

Le système de décodage que nous avons développé commence par segmenter et étiqueter en classes phonétiques (voyelles, fricatives, occlusives, sonantes et autres sons) la phrase grâce au module de segmentation automatique. La description acoustique des instances de triplets est ensuite construite automatiquement en faisant appel au suivi automatique de formants et au détecteur-analyseur de barres d'explosion. Comme les autres détecteurs acoustiques n'ont pas encore été réalisés notre système se limite à identifier les triplets centrés sur une voyelle ou une occlusive.

Afin de vérifier la pertinence de la description acoustique des triplets, nous avons conçu et testé les modules d'identification pour les voyelles et les occlusives. Ces modules évaluent une distance acoustique grossière entre l'instance à reconnaître et les triplets de référence.

2.3.1 Triplets centrés sur une voyelle

La distance calculée ne fait intervenir, pour l'instant, que les trajectoires formantiques, après que le triplet inconnu et le triplet de référence ont été normalisés en temps. Le calcul consiste à évaluer le coût de déplacement des arcs de trajectoires formantiques du triplet inconnu vers ceux du triplet de référence, et pénalise lourdement les grands écarts de pentes formantiques. Pour chaque instance le module d'identification acoustique donne les triplets de référence les plus proches du triplet inconnu.

Nous avons testé ce module sur 25 voyelles extraites de deux phrases du corpus de parole continue « La bise et le soleil... » pour 10 locuteurs (soit 250 voyelles). Les voyelles de ces deux phrases sont bien réparties dans le triangle vocalique. Comme le nombre des triplets de ce test est limité nous avons seulement pris en compte les trajectoires formantiques du son central pour le calcul de la distance acoustique. Cela pénalise donc les résultats de notre module d'identification acoustique mais donne une idée plus juste des performances réelles de notre approche. En conservant les trois meilleurs candidats pour chaque voyelle nous avons obtenu 91% de bonnes reconaissances, ce qui est tout à fait encourageant. Après examen des erreurs, il apparaît qu'il s'agit dans la plupart des cas d'erreur de suivi de formants, ce qui souligne l'importance de disposer de détecteurs acoustiques très performants.

2.3.2 Triplets centrés sur une occlusive

En ce qui concerne les occlusives il faut ajouter au terme de distance concernant les trajectoires formantiques un terme destiné à évaluer la distance entre

deux barres d'explosion. Cette distance est calculée en considérant les concentrations d'énergie des barres d'explosion à comparer et consiste à trouver le plus grand recouvrement possible entre les deux bursts. Cette distance est pondérée par les coefficients de diffusion et de compacité. Le module d'identification acoustique correspondant a été testé sur 15 occlusives pour 10 locuteurs du même corpus que pour les voyelles. Les résultats en ne considérant que les deux meilleurs candidats sont moins bons (71%) que pour les voyelles. Pour améliorer ce taux de reconaissance nous étudions des procédures de comparaison de barres d'explosion plus évoluées qui séparent l'effet de la voyelle de celui de l'occlusive.

3 Techniques d'intelligence artificielle pour un système de décodage à base de triplets

Le choix du triplet phonétique répond aussi aux exigences suivantes : représenter et utiliser les conaissances acoustico-phonétiques d'une manière aussi souple et efficace que possible au cours du décodage. À ce titre l'avantage du triplet est d'être une unité d'information complète qui se prête bien aux techniques de l'intelligence artificielle comme nous allons le voir maintenant.

On peut déduire à partir de deux triplets de la base de conaissances un certain nombre de relations portant sur les positions relatives des formants décrivant ces triplets. Lors du décodage il est donc possible de vérifier que ces relations sont satisfaites entre deux instances de ces mêmes triplets, même s'ils sont très éloignés dans le temps. Pour assurer que le résultat du décodage est cohérent au niveau de la phrase, et donc que les relations précédentes sont satisfaites, nous avons utilisé un algorithme de relaxation flou [Mohr 86] qui élimine les étiquettes incohérentes du treillis phonétique. Avec des paramètres de relaxation appropriés 70% des étiquettes incorrectes sont éliminées. Ces premiers résultats sont encourageants mais montrent qu'il faut améliorer la construction des relations acoustiques (jouant le rôle de contraintes pour la relaxation) car certaines d'étiquettes correctes sont éliminées (environ 30%).

4 Perspectives

L'utilisation de triplets en décodage acoustico-phonétique soulève un certain nombre de sujets de recherche, parmi lesquels :

- **La modélisation des déformations que peut subir un triplet** Le choix du triplet comme unité phonétique repose sur l'idée que le triplet permet de « capturer » une bonne partie des effets de coarticulation et donc que sa réalisation acoustique varie peu. Malgré tout, il reste que le triplet est soumis à un certain nombre d'influences qui peuvent altérer sa réalisation. Certains effets de coarticulation ne peuvent pas être pris en compte, c'est le cas des séquences V_1CV_2 avec un consonne labiale où la première voyelle peut influencer la seconde. Par ailleurs, nous ignorons encore les effets de l'accentuation et de la position du triplet dans le mot et dans la phrase.

- **Détection des événements acoustiques** C'est sans doute le point qui, à l'heure actuelle, limite le plus les systèmes de décodage à base de conaissances. Le suivi de formant automatique est un problème classique pour lequel de nombreuses solutions ont été proposées. En revanche la détection des barres d'explosion et la détermination de ses caractéristiques est un problème peu souvent abordé et pour lequel il reste donc de nombreux progrès à accomplir. L'état des recherches est encore moins avancé en ce qui concerne l'étude fine des bruits de friction (comment trouver, par exemple, la limite inférieure de bruit en tenant compte des formants de bruit ?) et ce point nécessite donc encore de substanciels efforts.

- **combinaison des résultats partiels** Lors du décodage il faut combiner un certain nombre de résultats de natures différentes : segmentation, événements et indices acoustiques, déductions phonétiques, déductions lexicales... Traditionnellement on attache à chacun de ces résultats un score indiquant la confiance qu'on lui accorde. Les scores sont ensuite judicieusement combinés pour évaluer le niveau de confiance d'une déduction obtenue à partir de résultats partiels. Cependant ce mécanisme de contrôle ne permet pas de connaître l'ensemble des faits de base et des déductions partielles qui ont contribué à une déduction. Il est donc possible qu'il existe des incohérences entre plusieurs de ces faits et déductions : (*i*) un ou plusieurs événements acoustiques ont été interprétés de manière contradictoire (*ii*) plusieurs faits ou déductions partielles apparaissant à des niveaux d'analyse différents (segmentation, formants, indices acoustiques...) peuvent être incompatibles à la lumière de nos conaissances des phénomènes de la parole. Le raisonnement hypothétique [de Kleer 86] qui permet d'associer à une déduction l'ensemble des faits qui la sous-tendent et d'assurer qu'ils sont compatibles entre eux, représente une voie de recherches intéressante pour la parole. Dans ce cadre, il faut aborder deux points clé : d'une part, comment mettre en évidence les incohérences parmi un ensemble d'événements acoustiques et de déductions partielles ; d'autre part, comment traiter le cas où une incohérence apparaît?

De nombreux efforts restent donc à fournir dans le cadre d'une approche à base de triplets du décodage acoustico-phonétique, tant au niveau de la détection fine des événements acoustiques, qu'au niveau des techniques mises en œuvre pour utiliser le plus efficacement possible les conaissances acoustiques et phonétiques.

Références

[de Kleer 86] J. de Kleer. An assumption-based TMS. *Artificial Intelligence*, (28):127–162, 1986.

[Fohr 89] D. Fohr, N. Carbonell, and J.P. Haton. Phonetic decoding of continuous speech with the aphodex expert system. In *Proceedings of European Conference on Speech Technology*, pages 609–612, Paris, France, September, 1989.

[Laprie 90] Y. Laprie. Optimum spectral peak track interpretation in terms of formants. In *Proceedings of International Conference on Spoken Language Processing*, volume 2, pages 1261–1264, Kobe, Japan, November, 1990.

[Mohr 86] R. Mohr and C. Henderson. Arc and path consistency revisited. *Artificial Intelligence*, (28):225–233, 1986.