Segmentation du bruit d'explosion des occlusives

Yves Laprie et Anne Bonneau

LORIA/INRIA & CNRS
615 rue du jardin botanique 54602 Villers-lès-Nancy FRANCE
Mél: {Yves.Laprie, Anne.Bonneau}@loria.fr

Résumé

This paper investigates burst segmentation for the evaluation of acoustic cues used to identify unvoiced French stops. Unlike other works which utilize a fixed length window, our approach consists in segmenting bursts into transient and frication noise. The transient is found by minimizing the sum of spectral variances of transient and frication noise over the burst. The spectral variance criterion has the advantage of being sensitive both to energy deviations and spectral variations. Additional correction procedures augment the robustness of the segmentation against the presence of spurious noises during the closure and the determination of the voicing onset with delay. The relevance of our segmentation method has been evaluated by comparing the characteristics of the main spectral peak (energy prominence versus frequency) in the transient segmented by our method with those of the full burst. Our experiments showed that bursts segmented by our method allow a better discrimination between the three places of articulation.

1. Introduction

La segmentation du bruit des occlusives a été réalisée dans le but de définir des indices acoustiques pour les consonnes occlusives du français. Nous nous intéressons particulièrement aux indices qui sont à la fois bien marqués sur le plan acoustique et très discriminants sur le plan phonétique. Ces indices peuvent être positifs, auquel cas ils permettent d'identifier un son avec certitude, ou négatifs, auquel cas ils permettent d'éliminer un son avec certitude. Nos indices reposent sur les trajectoires formantiques et sur les caractéristiques spectrales des bruits.

Le bruit des occlusives peut être décomposé en trois segments [3], l'explosion, l'aspiration (absente en général en français) et le bruit de friction. Les transitions formantiques sont souvent visibles dans le bruit de friction et dans l'aspiration. Le bruit commence au relâchement de l'articulation (à la fin de la tenue) et, dans des séquences CV, se termine avec la première période vocalique. Ce bruit ne correspond pas à une articulation tenue, -il diffère en cela du bruit des fricatives-, mais au passage de l'articulation de la consonne à celle du son suivant, une voyelle en général. Par conséquent, le spectre du bruit change rapidement en fonction du temps et, dans la plupart des cas, est plus discriminant dans sa partie initiale. Au contraire, plus le spectre est calculé près de la voyelle suivante, plus il est dominé par les formants de cette voyelle. Ceci est particulièrement visible dans les spectres de Krull correspondant à des occlusives voisées[4] et calculés à deux instants différents (t1 : de 0 to 10 ms après le relâchement de la consonne et t2: de 10 to 20 ms après celui-ci).

Il semble donc judicieux d'évaluer les indices de lieu d'articulation des occlusives juste après le relâchement de l'articulation consonantique. La segmentation du bruit étant très difficile à réaliser, seul un segment de longueur fixe au début du bruit est généralement pris en compte (26 ms pour Stevens et Blumstein [1], ou 10 to 15 ms pour Zue [5]).

Nous avons donc segmenté le bruit en deux ou trois segments. Dans ce papier, nous présentons d'abord notre méthode de segmentation, puis nous comparons l'efficacité des indices obtenus avec ou sans segmentation.

2. Segmentation du bruit d'explosion

L'objectif de la segmentation est de trouver l'explosion. Pratiquement, cet objectif consiste à décomposer le bruit d'explosion en deux parties homogènes. Ces deux objectifs ne sont pas strictement équivalents car la décomposition repose sur l'optimisation d'un critère qui traduit l'objectif initial de manière approximative.

La toute première étape consiste à localiser le bruit d'explosion ce qui oblige à reconsidérer la structure du bruit d'explosion donnée plus haut. Dans le cas de bruits d'explosion multiples, en particulier pour le son /k/ dans le contexte d'une voyelle centrale, il n'y a plus deux mais souvent trois parties qui correspondent aux événements suivants : **(i)** la première explosion suivie d'un silence, **(ii)** les autres explosions qui sont regroupées et forment l'explosion principale, **(iii)** le bruit de friction. Cette décomposition en trois parties est nécessaire quand le silence qui sépare la première explosion des suivantes ne peut pas être négligé, c'est-à-dire quand sa durée excède 10 ms. Elle influence la détection du début du bruit d'explosion, comme nous le verrons plus bas.

La fin du bruit d'explosion est donnée par le début du voisement déterminé à l'aide de l'algorithme de calcul de la fréquence fondamentale que nous avons développé précédemment. Tous les calculs de localisation et de décomposition ont lieu à partir d'un spectro-

gramme à bande large avec une fenêtre de 4 ms et un déplacement de 1 ms. Ces choix conduisent à une précision de 1 ms. Le début du bruit d'explosion correspond à l'apparition d'une énergie sonore significative. Le seuil de détection est relativement faible mais nous imposons l'apparition d'un pic d'au moins 6 dB au-dessus du niveau du silence utilisé pour l'affichage du spectrogramme. La première explosion d'une explosion multiple est cherchée sous la forme d'un pic intense et brutal suivi d'un silence court, puis d'un pic plus intense que le premier. Nous imposons que la première explosion ne soit pas plus intense que l'explosion principale.

2.1. Quel critère de segmentation ?

L'observation d'un grand nombre de bruits d'explosion nous a conduit à distinguer plusieurs grandes classes de bruit d'explosion qui peuvent être utiles pour guider le développement d'un algorithme de segmentation. La Fig. 1 donne la représentation schématique de ces grandes classes avec pour chacune d'elle un exemple et le profil de la courbe d'énergie.

Les deux classes susceptibles d'être concernées par la définition d'indices acoustiques forts sont les deux premières. Les bruits d'explosion de la première classe sont caractérisées par des explosions très intenses qui dominent le reste du bruit d'explosion. Par ailleurs, les spectres de l'explosion et du bruit de friction sont sensiblement différents. Les bruits d'explosion moyens (ceux de la deuxième classe) présentent des explosions dont le spectre diffère nettement de celui du bruit de friction mais dont l'énergie de l'explosion ne domine pas nettement celle du bruit de friction. Les explosions de la dernière classe ne portent que peu d'information facilement utilisable car leur énergie est trop faible ce qui rend la segmentation en deux parties hasardeuse. Il faut noter que cela ne signifie pas que l'explosion n'existe pas mais que sa faible énergie la rend difficile à segmenter et qu'il n'est d'ailleurs pas certain que ses caractéristiques spectrales soient perçues par le système auditif périphérique.

Nous avons ainsi expérimenté deux critères qui font appel au calcul de spectres à bande large. Nous notons $X(e^{j\omega_k})$ la transformée de Fourier du signal, $S(e^{j\omega_k}) = max(20\,log_{10}|X(e^{j\omega_k})| - dB_{SILENCE}, 0)$, l'énergie au-dessus du seuil de silence utilisé dans le calcul du spectrogramme, et $\overline{S}(e^{j\omega_k})$ la valeur moyenne de $S(e^{j\omega_k})$ sur le segment de parole délimité par les instants t_0 et t_1. Le premier critère est la somme de la corrélation entre les spectres d'un segment de parole et leur spectre moyen :

$$Correl(t_0,t_1) = \sum_{t=t_0}^{t=t1} \frac{\sum_{k=K_0}^{k=K_1} S_t(e^{j\omega_k})\overline{S}(e^{j\omega_k})}{\sqrt{\sum_{k=K_0}^{k=K_1} S_t(e^{j\omega_k})^2}\sqrt{\sum_{k=K_0}^{k=K_1} \overline{S}(e^{j\omega_k})^2}}$$

où K_0 and K_1 représentent le domaine spectral à prendre en compte. La meilleure segmentation correspond à l'instant pour lequel la somme de l'homogénéité spectrale des deux segments est maximale. Le critère à maximiser est donc : $Correl(t_i, limit) + Correl(limit, t_f)$ où t_i et t_f sont les instants de début et de fin du bruit d'explosion, et $limit$ la frontière entre l'explosion et le bruit de friction. Le second critère est la variance spectrale calculée sur chaque segment :

$$VarSpec(t_0,t_1) = \frac{1}{t_1 - t_0} \sum_{t=t_0}^{t=t1} \sum_{k=K_0}^{k=K_1} (S_t(e^{j\omega_k}) - \overline{S}(e^{j\omega_k}))^2$$

La meilleure segmentation est l'instant pour lequel la somme des variances est minimale.

Le critère de corrélation favorise une segmentation en régions dont le spectre de parole a la même allure, donc a priori la même origine acoustique. Le second critère favorise l'émergence de régions dont l'énergie spectrale s'écarte peu du spectre moyen. Il est apparu que la segmentation découlant du critère de variance était plus conforme avec la segmentation que nous aurions faite manuellement. En particulier, elle donne de bons résultats aussi bien quand l'explosion est très intense que quand les spectres de l'explosion et du bruit de friction sont différents. Nous avons donc retenu le critère de variance.

2.2. Robustesse de la segmentation

Les erreurs de segmentation viennent soit de la structure acoustique du bruit d'explosion, soit du contexte dans lequel la segmentation est utilisée. Nous présentons maintenant les méthodes destinées à augmenter la robustesse de la segmentation vis-à-vis des erreurs de détermination du début du bruit d'explosion ou du début du voisement. En effet, l'algorithme de segmentation repose sur l'idée que le bruit d'explosion peut être décomposé en deux parties. Il risque donc de commettre des erreurs quand cette hypothèse n'est pas vérifiée, soit parce que le début du voisement est détecté trop tard, soit parce qu'un bruit parasite précède le bruit d'explosion.

La première difficulté est liée à l'imprécision de la détermination du début de voisement. La prise en compte d'une ou de plusieurs périodes de voisement, en plus du bruit d'explosion, peut en effet modifier la position de la frontière, car elle conduit à un faux bruit d'explosion avec trois parties spectralement et énergétiquement différentes : l'explosion, le bruit de friction et les premières périodes de voisement. Comme les régions extrêmes (l'explosion et les premières périodes de voisement) concentrent l'essentiel de l'énergie acoustique, la segmentation donne en général une frontière quelque part dans le bruit de friction. Nous avons réduit l'influence de l'énergie à la fin du bruit d'explosion pour éliminer ce risque d'erreur. Pratiquement, nous avons remplacé le terme d'écart dans le calcul de la variance par :

$(S_t(e^{j\omega_k}) - \overline{S}(e^{j\omega_k}))^2 \times cos^{\alpha}(\frac{(t-t_i)}{t_f - t_i}\pi/2)$ où t_i et t_f sont les extrémités du bruit d'explosion. Cela signifie que : $t_i \leq t_0 \leq t_1 \leq t_f$. L'exposant α est utilisé pour régler l'influence des derniers spectres et dépend donc de la précision de la détermination du voisement.

La seconde difficulté est liée à la présence de bruits parasites pendant la fermeture du conduit vocal qui provoquent de fausses alarmes. Nous avons choisi un seuil d'énergie faible pour pouvoir trouver les attaques de bruits d'explosion faibles, par exemple celle du troisième bruit d'explosion de la Fig. 1. Schématiquement, ces fausses alarmes correspondent à la présence

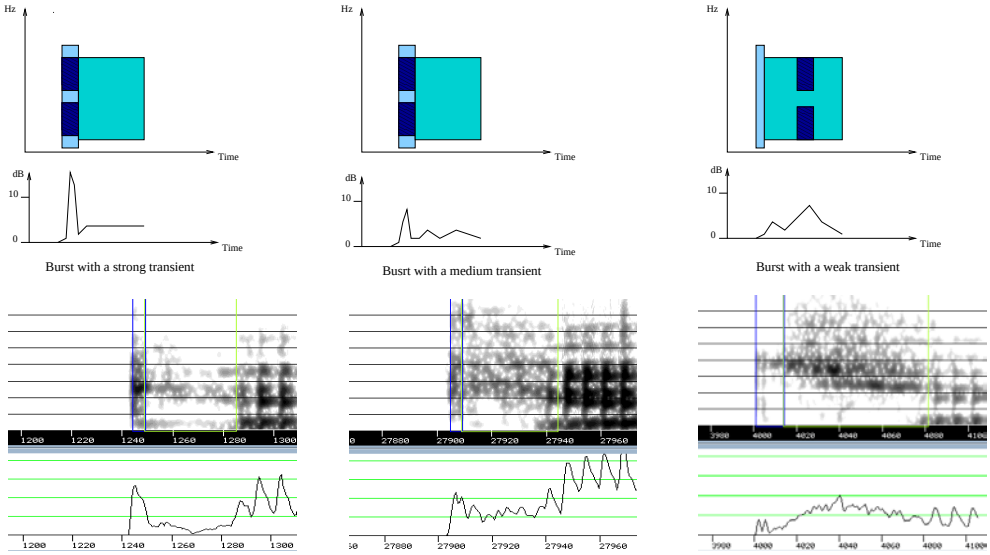

FIG. 1 – *Les trois grandes classes de bruit d'explosion. De haut en bas : la représentation schématique (les rectangles noirs représentent les proéminences spectrales de l'explosion, et les rectangles gris représentent le bruit de friction), le profil d'énergie type, le spectrogramme (une graduation par kHz sur l'axe des fréquences) d'un exemple réel et la courbe d'énergie(une graduation tous les 5 dB) correspondante. De gauche à droite : /ka/, /kœ/, /ty/. La segmentation automatique est superposée au spectrogramme.*

d'un petit bruit suivi d'un silence et du vrai bruit d'explosion. Il est possible que plusieurs bruits parasites précèdent le bruit d'explosion mais cela ne modifie pas la procédure de correction décrite plus bas. Le bruit parasite correspond toujours à l'explosion. Le silence qui suit le bruit peut être fusionné avec l'explosion ou le bruit de friction. La procédure de correction différencie ces deux cas.

Le bruit parasite et le silence forment la barre d'explosion Le spectre moyen de l'explosion est calculé après la segmentation. Si le silence est fusionné avec le bruit parasite le spectre moyen est très faible voire nul, et par conséquent, il ne peut pas s'agir de l'explosion. La procédure de segmentation est donc réitérée en déplaçant le début du bruit d'explosion jusqu'à ce que le spectre de l'explosion soit au moins 3 dB plus élevé que le seuil de silence pour le calcul du spectrogramme.

Le bruit parasite joue le rôle de l'explosion Cela signifie que le silence et le vrai bruit d'explosion sont fusionnées et qu'ils représentent à tord le bruit de friction. Le bruit de friction est donc segmenté à son tour après la première segmentation. S'il y a un silence d'une durée significative (plus de 20 ms) cela signifie que le bruit de friction est en fait formé d'un silence suivi du vrai bruit d'explosion qui est alors segmenté.

Cette double correction sur l'explosion et le bruit de friction élimine la plupart des erreurs dues à des bruits parasites, mais sans éliminer les explosions multiples qui correspondent à des pics d'énergie significativement plus intenses et suivis par un silence très court (moins de 10 ms).

3. Évaluation

Nous avons testé notre méthode de segmentation en comparant l'efficacité des indices calculés sur l'attaque par rapport à celui des indices calculés sur le bruit complet. Les principaux indices acoustiques fournis par le bruit sont la fréquence du pic le proéminent, désigné ci-dessous comme «le pic principal», et la répartition de l'énergie (compacte ou diffuse). Stevens et Blumstein[1] ont qualifié le spectre des alvéolaires de «diffuse-rising», celui des palato-vélaires de «compact», et celui des consonnes labiales de «diffuse-falling». En général, la fréquence du pic principal est située au-dessus de 2200-2500 Hz pour les dentales (si le pic correspondant au locus dental ne domine pas le spectre), en-dessous de cette même région fréquentielle pour les labiales et correspond au F2 ou au F3 de la voyelle suivante pour les vélaires. La variabilité intra- et inter-contextuelle des indices des occlusives est très importante (voir Fig. 2) et la fréquence du pic principal ne peut être fiable si celui-ci n'est pas suffisamment proéminent. Pour les labiales, le bruit est généralement trop faible pour permettre une identification fiable de cette consonne[5, 2]).

Les indices acoustiques sont d'autant plus efficaces que la proéminence est suffisamment forte et que la fréquence du pic est située dans la région attendue. Nous avons donc projeté les consonnes dans un plan formé par la fréquence du pic principal (en abcisse) et sa proéminence (en ordonnée). Nous avons testé la proéminence du pic principal par rapport à l'energie moyenne d'une part, et par rapport à l'énergie du deuxième pic le plus proéminent, d'autre part. Le second critère est apparu un peu plus efficace que le premier et nous l'avons retenu. Nous n'avons pas testé ici les critères de compacité, plus délicats à définir.

Afin de valider notre méthode de segmentation, nous avons comparé le pouvoir de discrimination du bruit complet avec celui de l'attaque, à l'aide des caractéristiques du pic principal (sa fréquence et sa proéminence). Le bruit complet va du relâchement de l'articulation jusqu'à la première période de la voyelle suivante. Nous avons effectué notre test sur une série de séquences /?-stop-V/ extraites de deux cor-

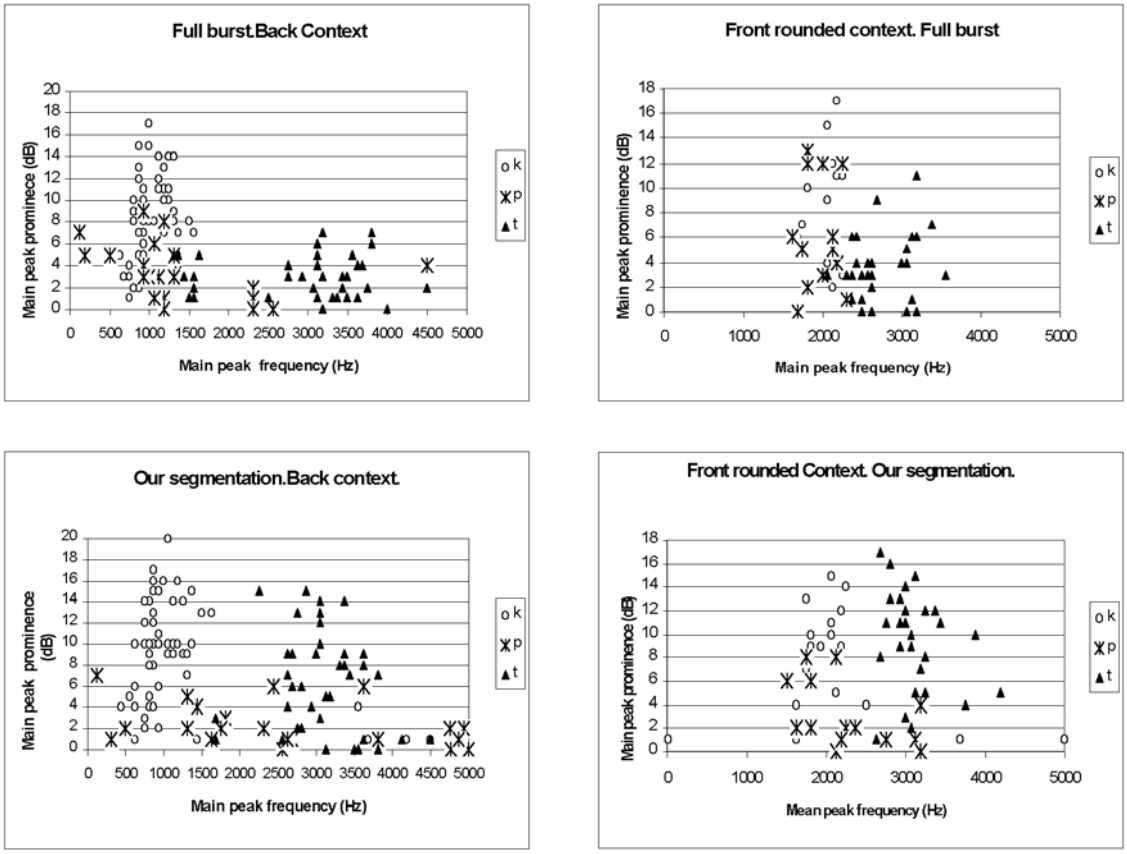

FIG. 2 – *Fréquence (en abcisse) et émergence (en ordonnée) du pic spectral le plus proéminent, pour deux contextes vocaliques. Comparaison des résultats pour le bruit complet et pour l'attaque, segmentée selon nos critères.*

pora français. Le premier corpus (un sous-ensemble de BDSONS) contient des mots isolés prononcés par 5 locuteurs masculins. Le deuxième corpus contient 22 phrases lues formées de consonnes occlusives et de voyelles et répétées trois fois par 4 locuteurs masculins. Nous avons analysé approximativement 500 consonnes. Nous commenterons les résultats pour deux contextes vocaliques en détail, les voyelles d'avant arrondies et les voyelles d'arrière.

4. Discussion

Pour les voyelles antérieures arrondies, les résultats montrent clairement que les pics principaux sont plus concentrés dans la région du F2-F3 de la voyelle suivante (environ 1700-2400 Hz) quand les indices sont calculés pour le bruit qu'ils ne le sont à l'attaque. Cette concentration est essentiellement due à l'abaissement fréquentiel du pic de certaines labiales et de la plupart des dentales. Pour le bruit complet, nous pouvons également constaté une baisse drastique de l'énergie du pic principal, de telle sorte que, même quand les dentales ont une fréquence élevée, elles ne peuvent être identifiées de manière fiable. Enfin, l'augmentation de la proéminence de certains pics labiaux combinée avec la baisse de la proéminence des pics principaux des vélaires, réduit le nombre d'évidences claires de cette dernière classe de consonnes.

Les mêmes grandes tendances peuvent être observées pour le contexte arrière. L'abaissement du pic principal de certaines dentales et la plus grande concentration des labiales dans la région du maximum vélaire réduit le pouvoir de discrimination de la fréquence en tant qu'indice pour l'identification du lieu des occlusives quand on considère le bruit complet. Cette réduction est amplifiée par la baisse drastique de la proéminence du pic principal des dentales et des vélaires. Ces exemples montrent que l'attaque, convenablement segmentée par notre méthode, fournit des indices plus discriminants que le bruit complet. Cette segmentation peut être exploitée pour renforcer spécifiquement certains indices acoustiques afin d'améliorer l'intelligibilité de la parole, améliorer la détermination des paramètres d'un synthétiseur à formants afin de générer des stimuli proches de la parole naturelle et dans le cadre de la reconnaissance automatique de la parole.

Références

[1] S. E. Blumstein and K. N. Stevens. Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. *J. Acoust. Soc. Amer.*, 66:1001–1017, 1979.

[2] T. J. Edwards. Multiple features analysis of intervocalic English plosives. *J. Acoust. Soc. Am.*, 69(2):535–547, 1981.

[3] G. Fant. Stops in CV syllables. In *Speech sounds and features*. The MIT Press, Cambridge, 1973.

[4] D. Krull. Relating acoustic properties to perceptual responses: a study of Swedish voiced stops. *J. Acoust. Soc. Am.*, 88(6):2557–2570, 1990.

[5] V. W. Zue. *Acoustic characteristics of stop consonants: a controlled study*. PhD thesis, MIT, Lincoln Laboratory, 1976.