Reconnaissance robuste de parole en environnement réel à l'aide d'un réseau de microphones à formation de voie adaptative basée sur un critère des N-best Vraisemblances Maximales

L. Brayda[1,2]*, C. Wellekens*[1]*, M. Omologo*[2]

[1]Institut Eurecom
2229 Route des Cretes , 06904 Sophia Antipolis, France
Mél : brayda,welleken@eurecom.fr - http ://www.eurecom.fr/ brayda

[2]ITC-irst
Via Sommarive 18, 38050 Povo (TN), Italy
omologo@itc.it

ABSTRACT

Distant-talking speech recognition in noisy environnements is generally tackled by using a microphone array and a related multi-channel processing. Based on that framework, this paper proposes an *N-best* extension of the Limabeam algorithm, that is an adaptive maximum likelihood beamformer. *N-best* hypothesized transcriptions are generated at a first recognition step and then optimized independently one to each other. As a result, the *N-best* list is re-ranked, which allows selection of the maximally likely transcription to clean speech models. Results on real data show improvements over both Delay and Sum Beamforming and Unsupervised Limabeam at low SNR and with moderate reverberation.

1. INTRODUCTION

Les perfomances des systèmes de reconnaissance de la parole baissent nettement dans un environnement réel et d'autant plus que le locuteur se trouve loin du microphone dans un bruit soit additif, soit convolutif. Des études précédentes [1] ont montré que la qualité du signal vocal peut être améliorée (augmentation du rapport signal à bruit) par l'utilisation des réseaux de microphones. En exploitant la corrélation spatiale entre les signaux multi-canaux, on peut focaliser le réseau vers le locuteur (formation de voie ou *beamforming*). Ceci peut se faire soit en exploitant simplement l'interférence destructive du bruit par une technique de retard-et-sommation (R&S) [2], soit en appliquant des filtres à chaque canal (filtrage-et-sommation). Ces filtres peuvent être fixes ou adaptés pour chaque échantillon ou trame selon un certain critère [3, 4]. Le problème qui peut se poser est que l'amélioration du rapport signal à bruit (RSB) n'entraine pas nécessairement une augmentation concommittante de la performance du reconnaisseur [5]. Seltzer [6, 7] propose d' appliquer une technique de filtrage adaptatif sur les signaux multicanaux sous un critère de Vraisemblance Maximale (Limabeam) et non plus de rapport signal à bruit. Dans cette méthode les filtres sont adaptés de façon aveugle en utilisant les parametres des modèles acoustiques non-bruités qui alignent le mieu les vecteurs acoustiques bruités. Le reconnaisseur utilisera ensuite la somme des signaux filtrés pour générer une transcription finale. Dans une étude récente [8] nous avons montré que si l'on considère en parallèle les *N-best* hypothèses au lieu de la meilleure hypothèse pour adapter les filtres, on peut augmenter les performances du reconnaisseur et approcher celles d'un algorithme supervisé. Dans ce papier nous testons cette méthode améliorée dans un environnement réel et nous montrons que les performances du Limabeam peuvent être encore augmentées.

2. L'ALGORITHME LIMABEAM

L'algorithme Limabeam utilise un réseau de L microphones auquel on applique une formation de voie de type filtrage et sommation. Les coefficients des filtres non récursifs (RIF) de degré M, un par microphone, sont modifiés de façon adaptative. Un tel formateur de voie peut être représenté comme suit :

$$x[k] = \sum_{m=1}^{M} h_m[k] * s_m[k] \tag{1}$$

où $s_m[k]$ est le signal discret dans le domaine temporel reçu au m-ème microphone, $h_m[k]$ est la réponse impulsionnelle du filtre RIF du m-ème canal, $x[k]$ est la sortie du formateur, * dénote la convolution et k est l'index temporel. L'ensemble des filtres peut être représenté par un super-vecteur $\mathbf{h}$. Pour chaque trame, les composantes du vecteur acoustique sont calculées et exprimées en fonction de $\mathbf{h}$:

$$\mathbf{y}_L(\mathbf{h}) = \log_{10}\left(W\,|\,\mathrm{FFT}(\mathbf{x}(\mathbf{h}))|^2\right) \tag{2}$$

où $\mathbf{x}(\mathbf{h})$ est le vecteur observé, $|\,\mathrm{FFT}(\mathbf{x}(\mathbf{h}))|^2$ est le vecteur des différents composants du spectre de puissance, W est la matrice des coefficients Mel et $\mathbf{y}_L(\mathbf{h})$ est le vecteur des log-énergies des bancs de filtre (LFBE). Les coefficients cepstraux sont dérivés par une transformation en cosinus discrète (DCT).

$$\mathbf{y}_C(\mathbf{h}) = \mathrm{DCT}\left(\mathbf{y}_L(\mathbf{h})\right). \tag{3}$$

Limabeam vise à dériver un ensemble de *M* filtres RIF, qui maximisent la vraisemblance de $\mathbf{y}_L(\mathbf{h})$ étant donné un alignement Viterbi estimé d'une transcription supposée. Ceci est exprimé par :

$$\hat{\mathbf{h}} = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h})\,|\,w) \tag{4}$$

où w est la transcription supposée. L'optimisation est faite par gradient conjugué non linéaire. L'alignement Viterbi fait sur la sortie du réseau peut être estimé soit à partir de la transcription obtenue après une première étape de reconnaissance (Unsupervised Limabeam), soit en supposant que la transcription correcte est disponible (Oracle Limabeam). Plus de détails peuvent être trouvés dans [5]. L'Unsupervised Limabeam fonctionne bien dans les environnements bruyants, même avec un seul canal. Cependant, les expériences préliminaires conduites sur des données simulées [8] ont indiqué deux faits : d'abord, les résultats de l' Oracle Limabeam sur un canal simple étaient proches du R&S simple sur huit canaux ; en second lieu, il y avait toujours une marge d'amélioration possible entre l'Unsupervised et la version Oracle appliqués aux signaux multi-canaux.

3. APPROCHE *N-best* À OPTIMISATION PARALLÈLE

L'algorithme de Limabeam augmente la vraisemblance de l'hypothèse de la première transcription après une première étape de reconnaissance. Nous proposons d'appliquer *N-best* optimisations indépendantes et en parallèle : cette approche est basée sur le fait que la liste des *N-best*, avant l'optimisation parallèle, est triée par vraisemblance et pas nécessairement par le taux d'erreur en mots (WER), qui devrait être le critère optimal. En appliquant l'algorithme Limabeam sur chaque hypothèse, le tri de la liste des *N-best* hypothèses change parce que les hypothèses à WER inférieurs sont mieux optimisées, même s' ils ont une vraisemblance initiale inférieure. Nous prouvons au niveau expérimental que la nouvelle hypothèse choisie (la nouvelle hypothèse à vraisemblance maximale) dans cette nouvelle liste a, en moyenne, un WER inférieur à la première choisie dans la liste précédente (voir Figure 1)

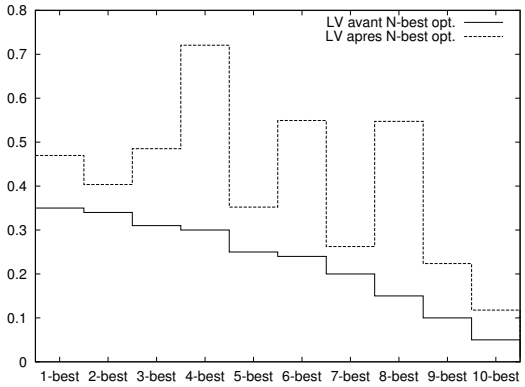

FIG. 1: *Exemple de Log-vraisemblance normalisée d'une phrase dont les 10 meilleures hypothèses sont optimisées. Avant l'optimisation les phrases sont triées par vraisemblance. Après, les vraisemblances de toutes les hypothèses sont augmentées et l'hypothèse 4, qui a un WER inférieur à l'hypothèse 1, est maintenant la meilleure.*

Il est à noter qu'ici "*N-best*" résulte d'une réduction préliminaire à une liste qui n'inclut pas de répétitions de la même phrase, qui pourraient résulter d'un nombre et de localisations différentes d'unités de bruit ou de silence. Le système est décrit ci-dessous. Pour chacune des *N-best* hypothèses nous dérivons un ensemble de filtres RIF :

$$\hat{\mathbf{h}}_n = \arg\max_{\mathbf{h}} P(\mathbf{y}_L(\mathbf{h}) \,|w_n\,) \tag{5}$$

où w_n est la transcription à la première étape de reconnaissance, $P(\mathbf{y}(\mathbf{h}) \,|w_n\,)$ est la vraisemblance de la phrase observée étant donnée la *n*-meilleure transcription considérée. Notez que l'équation (5) est équivalente à Unsupervised Limabeam quand *n* est 1. Après que tous les *N-best* vecteurs RIF aient été optimisés en parallèle, de nouveaux vecteurs acoustiques sont calculés et une deuxième étape de reconnaissance est exécutée. La transcription de vraisemblance maximale est alors choisie :

$$\hat{n} = \arg\max_{n} P(\mathbf{y}_C(\hat{\mathbf{h}}_n) \,|\hat{w}_n\,) \tag{6}$$

où $\hat{w}_n$ est la transcription produite à la deuxième étape de reconnaissance et $\hat{n}$ est l'index de la transcription la plus vraisemblable, soit $\hat{w}_{\hat{n}}$. L'optimisation est faite dans le domaine LFBE, alors que la reconnaissance est faite dans le

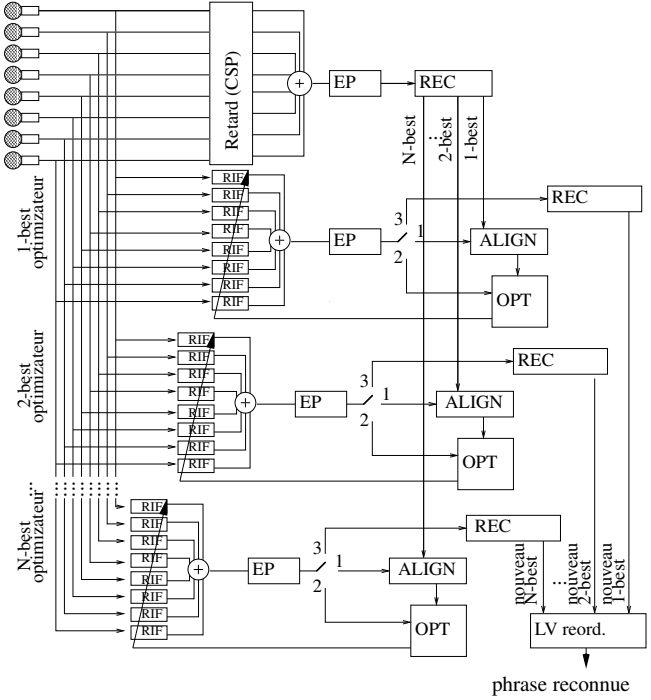

FIG. 2: *Schema du N-best Unsupervised Limabeam.*

domaine Cepstral comme dans [7]. Le nouvel ordonnancement des vraisemblances est de même fait dans le domaine Cepstral. Le système que nous proposons est représenté à la Figure 2. Le signal venant d'un réseau de microphones est traité par l'intermédiaire d'un R&S conventionnel, puis l'extraction des paramètres acoustiques (EP) et une première étape de reconnaissance est exécutée (REC). Le système de reconnaissance basé sur les modèles de Markov cachés (HMM) produit *N-best* hypothèses. Pour chaque hypothèse et en parallèle, l'algorithme de Limabeam est appliqué : d'abord un alignement Viterbi est effectué (commutateur sur 1 : ALIGN) et fixé, puis les coefficients des filtres RIF sont optimisés de manière adaptative en appliquant un algorithme de gradient conjugué (commutateur sur 2 : OPT). Un fois que la convergence est atteinte, les *N-best* séquences de vecteurs acoustiques sont identifiées (commutateur sur 3 : REC) et un ensemble différent de nouvelles transcriptions est produit. En conclusion, le dernier bloc compare les nouvelles *N-best* Log-Vraisemblances (LV-réordonnancement) en choisissant la plus élevée et la phrase reconnue est produite. Nos expériences montrent qu'en appliquant une approche N-best, l'Oracle Limabeam tel que proposé dans [7] ne constitue plus une limite supérieure à la performance du Limabeam : un alignement du type Baum Welch devrait produire une correspondance plus fine, avec en conséquence une meilleure optimisation. Pour obtenir une nouvelle borne, nous avons introduit la connaissance de la phrase correcte dans le bloc LV-reordonnacement : au lieu de (6), on choisit l'hypothèse dont la distance à la phrase connue est minimale. L'approche en aveugle *N-best* est donc couplée à une évaluation *a-posteriori* de la meilleure hypothèse : ceci est un indice de la qualité de la vraisemblance comme critère de choix.

4. BASE D'ÉVALUATION ET ENVIRONNEMENTS

Les expériences ont été conduites à l'aide du système de reconnaissance HTK basé sur les HMM, entraîné sur le corpus TI-digits. Les modèles de mots sont représentés

par des HMMs de type gauche-droite à 18 états, dont les distributions sont définies par une Gaussienne. La base d' éntrainement se compose de 8440 phrases, prononcées par 110 locuteurs (55 hommes et 55 femmes). La base de test se compose de 1001 phrases : elle a été enregistrée dans la salle décrite en Figure 3.

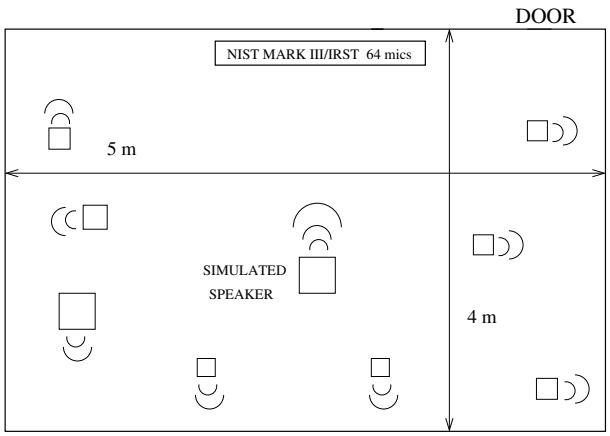

FIG. 3: *Chambre d'acquisition des données : les données non-bruitées sont émises par le haut-parleur central, le bruit simultanément par 8 sources. Le RSB à la source est de 0dB.*

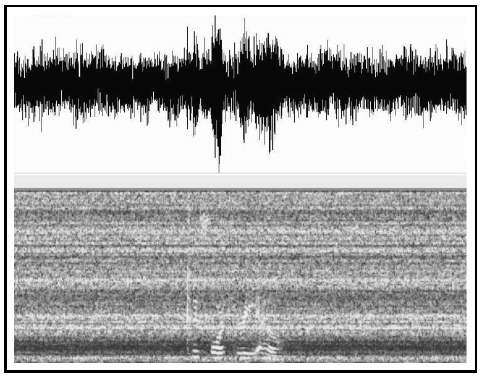

FIG. 4: *Spectrogramme d'une phrase enregistré avec un microphopne du MarkIII : le microphone capte les 8 bruits émis par les sources simultanées et distribuées et la phrase non-bruitée émise par le haut-parleur face au réseau.*

La salle mesure 5 x 4 mètres et elle présente un temps de propagation court (143 ms), ce qui nous a permis dans un environnement réel d'étudier les effets du bruit additif plutôt que ceux du bruit convolutif. Les données ont été produites par un haut-parleur de haute qualité (Tannoy 600A Nearfield Monitor). Le bruit, dont le spectre est visible dans la figure 4 a été engendré par 8 sources simultanées pour un rapport signal-bruit moyen de 0dB. Ce rapport est mesuré à la source et le vrai RSB mesuré au microphone varie selon la localisation des haut-parleurs et des microphones : on simule ainsi de meilleure façon un environnement réel. Les signaux ont été enregistrés par le réseau de microphones NIST MarkIII/IRST[9], placé à 1.3 mètres du haut-parleur Tannoy. Ce dernier est un réseau linéaire de 64 microphones, dont les capteurs sont espacés de 2 cm. Pour nos expériences nous avons choisi d' utiliser 8 microphones, espacés de 16 cm : cette configuration représente un bon compromis entre les hautes performances qui dépendent du nombre de microphones, le respect du théorème du recouvrement spatial, le besoin d'une complexité gérable et d'un temps de réponse raisonnable (pour l'optimisation des filtres).

Le MarkIII acquiert les données audio à une fréquence d'échantillonage de 44.1 kHz : dans cet environnement reel, nous avons constaté que les performances dépendent relativement peu de la fréquence d'échantillonage, et donc les données ont été sous-échantillonnées à 16 kHz avec un filtre polyphase à trois étapes. Les filtres RIF ont une longueur de 10 échantillons. L'extraction des paramètres acoustiques génère 12 coefficients Cepstraux en échelle Mel (MFCC) et la log-énergie ainsi que les premières et deuxièmes dérivées, pour un total de 39 coefficients. Les paramètres ont été calculés chaque 10 ms, en utilisant une fenêtre glissante de Hamming de 25 ms. La gamme de fréquences couverte par le banc de filtres a été limitée à 100-7500 Hertz pour éviter des bandes de fréquence où l'énergie de parole est limitée. La normalisation par la moyenne du cepstre est appliquée (CMN). Alors que la reconnaissance est exécutée dans le domaine cepstral, l'optimisation est faite dans le domaine LFBE en utilisant des vecteurs acoustiques d'ordre 16 et une distribution Gaussienne pour les modèles, mais sans CMN [7]. L'implémentation de l'algorithme Limabeam n'a nulle part été modifiée par rapport au travail original afin d'assurer la conformité aux expériences de Seltzer.

5. RÉSULTATS ET DISCUSSION

L'environnement choisi permet d'obtenir a-priori des hautes performances avec une technique R&S, qui marche d'autant mieux que le bruit additif est diffus. Ceci est évident en regardant les performances de chaque microphone (Tableau 1) et celles du R&S (première ligne du Tableau 2) : les microphones plus proches du haut-parleur

TAB. 1: *Performance du reconnaisseur sur chaque microphone choisi du MarkIII. Les meilleurs résultats sont observés là où le microphone est le plus proche du haut-parleur. Résultats fournis en précision de mots, c'est à dire en tenant compte des insertions.*

mic	1	9	17	25
Pre.	50.76%	57.26%	63.91%	61.46%
mic	33	41	49	57
Pre.	62.52%	64.21%	62.76%	52.69%

TAB. 2: *Performance des différents formateurs de voie : R&S, Unsupervised Limabeam (U.L.), N-best Limabeam (N-best L.), Oracle Limabeam (O.L.) et a-posteriori N-best Limabeam (a-post). L'optimisation considère jusqu'à 40 hypothèses en parallèle. Pour chaque méthode, on indique si l'optimisation est aveugle (AV) ou supervisée (SUP), son résultat en précision de mots (Pre) et son amélioration relative (AR) par rapport au R&S. Il est à noter que le a-posteriori N-best est une limite supérieure de reconnaissance par l'N-best Limabeam, parce qu'il optimise les RIF de façon aveugle, mais choisit, de façon supervisée, la phrase qui maximise la précision au lieu de celle qui maximise la vraisemblance.*

Mèthode	SUP	AV	Pre	AR
R&S	-	-	80.74%	-
U.L.		X	83.16%	12.5%
O.L.	X		83.49%	14.2%
N-best L. (40).		X	83.83%	16%
a-post (40)	X	X	85.13%	22.8%

ont les meilleures performances et l'absence d'une symétrie des résultats par rapport au centre du réseau (microphone 33) est une conséquence de la diffusion non-symétrique du son et du bruit additif et convolutif dans la salle. Pour appliquer le R&S, les retards appliqués aux signaux multi-canaux sont estimés avec l' information de la phase du spectre croisé (CSP) [10, 11]. La bonne performance du R&S (80.74%) est atteinte grâce à l' échantillonage spatial efficace du réseau.

La Figure 5 montre le comportement du *N-best* Limabeam en fonction de la longueur de la liste des *N-best* hypothèses. Le point de départ de la courbe (83.16%) correspond au Unsupervised Limabeam, quand une seule hypothèse est considérée. Les résultats s'améliorent d'autant plus que la liste est longue. Un phénomène apparemment surprenant est le fait que le *N-best* Limabeam se situe au dessus de l'Oracle Limabeam : comme prévu en Section 2, un alignement qui considère tous les chemins pourrait augmenter les performance de l'Oracle. La courbe semble présenter une asymptote au delà des 34-meilleures hypothèses et y atteint le maximum de 83.86%. Ceci est dû à la présence de bonnes hypothèses dans la partie inférieure de la liste *N-best* et indique que considérer plus d'hypothèses est la clé pour obtenir de meilleurs résultats. Pour des RSB plus élevés, l'asymptote des performances devrait être atteinte plus vite, c'est à dire en considérant moins d'hypothèses. Le comportement non-monotone, visible aussi dans les résultats rapportés en [8], est dû à l'inconsistance entre les critères du maximum de vraisemblance et du minimum WER parce que nous savons que choisir la transcription la plus vraisemblable dans le bloc d'ordonnancement (cfr. Figure 2) n'implique pas un choix du type WER minimum. Ceci n'est pas le cas si l'on observe le comportement *a-posteriori N-best* Limabeam, où la courbe est strictement monotone. Ceci car considérer plus d'hypothèses accroit nécessairement les chances de choisir l'hypothèse correcte lorsque la décision repose sur un critère WER qui ne peut pas s'appliquer car on ne connaît pas la phrase correcte d'avance. Les améliorations absolues et relatives sont rapportées à la Table 2 : l'utilisation de Limabeam est clairement justifiée et dans cet environnement, les performances de la méthode non-supervisée se rapprochent de l'Oracle. Comme on peut l'observer à la Figure 5, une approche *N-best* dépasse l'Oracle fournissant une amélioration relative de 16% par rapport au R&S lorque 40 hypotheses sont traitées en parallèle. Dans les mêmes conditions, le *N-best* Limabeam *a-posteriori* peut atteindre une amélioration relative de 22.8% : ceci signifie qu'en modifiant le critère de réordonnancement, on peut atteindre les performances d'un reconnaisseur *a-posteriori*. Une façon de réaliser cet objectif est de pondérer d'avantage les hypothèses dont le LV croît plus pendant le pas d'optimisation. Cette solution est encore à l'étude. En outre, au cours de ce travail nous avons amelioré les taux de reconnaissance dans un environnement de bruit diffus dans lequel une formation de voie adaptative apporterait un gain généralement inférieur par rapport au R&S que dans des environnements à bruit plus directif. Ceci nous encourage à explorer différents environnements bruités et réverbérents c'est à dire de nous rapprocher des conditions typiques d'une salle de réunion.

6. REMERCIEMENTS

La collection des données à eté partiellement supportée par le projet de recherche IST EU FP6 HIWIRE. L. Brayda voudrait remercier le MESR (Ministère de l'Enseignement Supérieur et de la Recherche - France) et Istituto Trentino di Cultura pour avoir supporté ce travail.

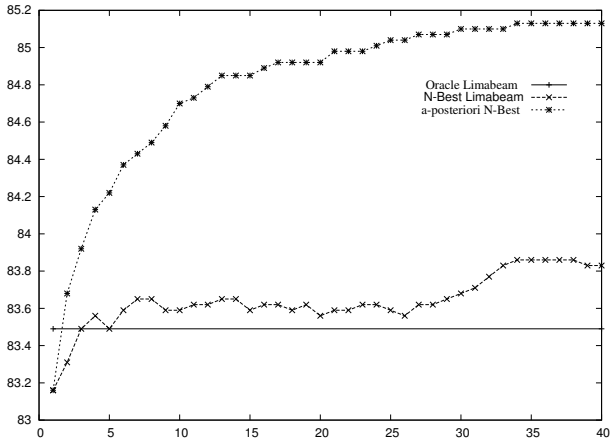

FIG. 5: *Performance de l' Oracle, du N-best Limabeam et du a-posteriori N-best Limabeam en fonction du nombre d'hypothèses considerées. Resultats exprimés en précision.*

RÉFÉRENCES

[1] M. Brandstein and D. Ward, *Microphone arrays - signal processing techniques and applications*, New York : Springer-Verlag, 2001.

[2] Johnson D and D. Dudgeon, *Array signal processing*, Prentice Hall, 1993.

[3] L. Griffith and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," in *IEEE Trans. on Antennas and Propagation*, 1982, vol. AP-30, pp. 27–34.

[4] O. Frost, "An algorithm for linearly constrained adaptive array processing," in *Proceedings of the IEEE*, 1972, vol. 60, pp. 926–935.

[5] Seltzer M., *Microphone array processing for robust speech recognition*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, USA, 2003.

[6] Seltzer M. and Raj B., "Speech recognizer-based filter optimization for microphone array processing," in *IEEE Signal Processing Letters*, March 2003, vol. 10, no, 3, pp. 69–71.

[7] Seltzer M., Raj B., and Stern R. M., "Likelihood-maximizing beamforming for robust hands-free speech recognition," in *IEEE Trans. on Speech and Audio Procesing*, September 2004, vol. 12, no, 5, pp. 489–498.

[8] Brayda L., Wellekens C., and Omologo M., "N-best parallel maximum likelihood beamformers for robust speech recognition," in *submitted to Proceedings of EUSIPCO*, Florence, Italy, 2006.

[9] Brayda L., Bertotti C., Cristoforetti L., Omologo M., and Svaizer P., "Modifications on NIST MarkIII array to improve coherence properties among input signals," in *AES, 118th Audio Engineering Society Convention, Barcelona, Spain*, 2005.

[10] M. Omologo and P. Svaizer, "Acoustic event localization using a cross-power spectrum phase based technique," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1994.

[11] C.H. Knapp and G.C. Carter, "The generalized correlation method for estimation of time delay," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1976, vol. 24, no, 4, pp. 320–327.