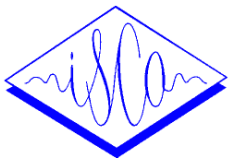

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

TALKER LOCALIZATION AND SPEECH RECOGNITION USING A MICROPHONE ARRAY AND A CROSS-POWERSPECTRUM PHASE ANALYSIS

D. Giuliani and M. Omologo and P. Svaizer

IRST-Istituto per la Ricerca Scientifica e Tecnologica,
I-38050 Povo di Trento (Italy)

ABSTRACT

Mismatch in training and testing conditions reduces considerably the performance of a speaker-independent HMM-based continuous speech recognizer. Compensation of this mismatch can avoid the complex and time-consuming retraining of the recognizer. This paper describes an acquisition system based on a four omnidirectional microphone array that was employed to reproduce a "beamformed" version of the original acoustic messages acquired in a noisy and reverberant environment, with a talker-microphone distance of one meter. In this preliminary activity, some simple noise compensation techniques (i.e. a Mean Spectrum based Enhancement and a Cepstrum Mean Subtraction) were incorporated in this preprocessing stage to obtain an enhanced version of the given utterance. Feeding a clean-condition trained continuous speech recognizer with enhanced signals led to a significant improvement of performance, if compared to the use of unprocessed single-microphone signals as input.

I. INTRODUCTION

Performance of a speech recognizer is often degraded drastically when the acoustic environment has different characteristics, with respect to the one for which it was designed, causing a mismatch in training and testing conditions [1, 2]. One reason for this is environmental noise. Another reason can be inconsistent use of the acoustic transducers during message acquisition. For example, a substantial decrease in performance can be caused by a non adequate position of the talker's mouth with respect to the microphone. Head-mounted, hand-held or fixed-position microphones do not solve this problem completely.

In the last decade, some research effort was devoted to microphone array processing techniques, especially for teleconferencing and large room recording, but also for speech recognition purposes [3]. Microphone arrays represent an attractive solution to the need for normalizing the speech information with respect to the above mentioned factors. Furthermore, using this technology allows investigation of ambitious tasks such as talker localization and talker tracking [4], that can imply new scenarios for future speech applications. The objective of this paper is to describe and to discuss the activities on this issue, recently conducted at our laboratories.

A system for acoustic event detection and localization in a noisy environment was developed and used for surveillance purposes [5]. The system runs in real time on a DSP-board (AT&T Surf-board) and allows acquisition of a generic message as well as acoustic source localization. The present version of the system includes an acquisition module connected to a 4-omnidirectional microphone array working at a sampling frequency of 48 kHz. The detection and localization processing is based on the use of a Crosspower-Spectrum Phase (CSP) technique, that provides high accuracy performance [5]. Once source position is estimated, a "delay and add" beamforming technique is applied to the given signals in order to reproduce an enhanced version (downsampled to 16 kHz) of the original acoustic message [6]. This enhanced version can still contain undesired components, generally due to the diffused noise present in the room; on the other hand, reverberation components are reduced. At this moment, our research efforts are oriented toward the application of this acquisition and preprocessing module as front-end of a continuous speech recognizer, trained with material acquired in a quiet environment. Here, the recognition system is based on the use of speaker independent continuous density HMM technology for large vocabularies: a description of its most recent version can be found in [7]. Retraining this speech recognizer, for given noisy conditions, could be a time-consuming procedure. Further, it would not solve the problem of mismatch between training and testing conditions, when the ambient noise characteristics change during the use of the recognizer. A first issue that was considered to improve noisy speech recognition performance concerned the combination of the CSP-based Time Delay Estimation(TDE) used for localization purposes with a "delay and add" beamformer and two successive spectral enhancement techniques, namely a Mean Spectrum based Enhancement (MSE) and a Cepstrum Mean Subtraction (CMS). To evaluate system performance, a database of speech sentences was collected with different environmental noise conditions and "talker-microphone array" distances.

The remainder of the paper is organized as follows. First, the talker localization system based on a microphone array and a CSP technique is described in Section II. Next, the enhancement techniques used in the acoustic preprocessing are discussed in Section III. The baseline of a speaker independent continuous speech recognizer is briefly presented in Section IV. Recognition performance comparison, using different enhancement preprocessing steps, is given in Section V. Finally, the objectives of our future work are described in Section VI.

II. TALKER LOCALIZATION

2.1 Problem Statement

The talker tracking problem, in a simplified approach, can be reconducted to the estimation of the acoustic

source position in a 2-dimensional space; for this purpose, at least three microphones must be employed, even if a higher number of microphones allows better accuracy. Following, we will always refer our discussion to the use of a linear array consisting of four omnidirectional microphones.

From a theoretical point of view, the signals acquired by each microphone can be assumed to be delayed replicas of the source signal plus noise: localizing the sound source implies the estimation of the time delays between the signals received. Once the delays are known, the acoustic event direction (and position) can be derived using geometry.

Now, let us assume that an acoustic source (e.g. a talker) located in position (x_s, y_s) (see Figure 1) generates an acoustic event $r(t)$ that is acquired by microphones $0, ..., (M-1)$ as signals $s_0(t), ..., s_{M-1}(t)$.

For the given source signal $r(t)$, propagated in a generic noisy environment, the signal acquired by the acoustic sensor i can be expressed as follows:

$$s_i(t) = \alpha_i r(t - \tau_i) + n_i(t) \qquad (1)$$

where α_i is an attenuation factor due to propagation effects, τ_i is the propagation time and $n_i(t)$ includes all the undesired components, which are assumed to be uncorrelated with $r(t)$.

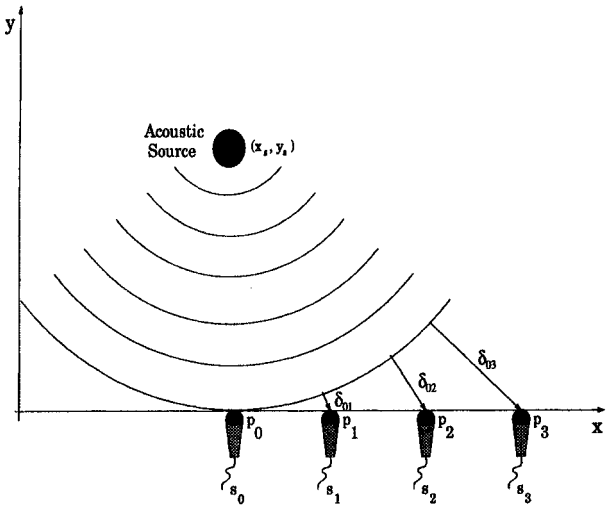

Figure 1: *Wavefront propagation of an acoustic stimulus generated in position* (x_s, y_s). *Signals* s_0, s_1, s_2, s_3 *are acquired through an array of microphones placed in positions* p_0, p_1, p_2, p_3. *The wavefront reaches microphones* $1, 2, 3$ *with delays* δ_{01}, δ_{02}, δ_{03}, *with respect to microphone* 0.

The relative delay of the wavefront arrival between the same microphones can be expressed as:

$$\delta_{ik} = (\tau_k - \tau_i). \qquad (2)$$

Different techniques can be exploited for Time Delay Estimation (TDE). In this work, the estimation of the sound source position is performed introducing information on delays in a representation called Coherence Measure (CM) [6], derived by a CSP analysis. CM can be suitable both when the acoustic source position does not change in time (stationary source), and when it changes (moving source), leading to a consistent approach when accurate talker tracking is required. In practice, a function $C_{ik}(t, \tau)$ is computed that represents the similarity between segments (centered at the time instant t) extracted from two generic signals s_i and s_k: this function is expected to have a prominent peak at the delay $\tau = \delta_{ik}$, corresponding to the direction of wavefront arrival.

2.2 CrosspowerSpectrum Phase

The most common method of determining the time delay δ_{ik} and the corresponding arrival angle θ_{ik}, given two signals $s_i(t)$ and $s_k(t)$, is to search for the lag τ which maximizes a cross-correlation function between them.

The CrosspowerSpectrum Phase (CSP) analysis derives from a mathematical modeling technique where signals are prefiltered before computing the correlation, leading to the so-called Generalized Cross-Correlation method. A detailed description of this theory can be found in [6, 8].

Given the signals s_i and s_k, at time t, the procedure for estimating the generalized correlation starts from the computation of spectra $\widehat{S}_i(t, f)$ and $\widehat{S}_k(t, f)$ through Fourier transforms applied to windowed segments of s_i and s_k, centered around time t.

The estimation of the normalized CrosspowerSpectrum, that can be expressed as follows:

$$\phi_{ik}(t, f) = \frac{\widehat{S}_i(t, f)\widehat{S}_k^*(t, f)}{|\widehat{S}_i(t, f)||\widehat{S}_k(t, f)|} \qquad (3)$$

preserves only information about phase differences between s_i and s_k. Then, the Coherence Measure is computed, that is the inverse Fourier transform $C_{ik}(t, \tau)$ of $\phi_{ik}(t, f)$. It is worth noting that, also in a real situation, the resulting function (defined in the lag axis τ) has a constant energy, mainly concentrated on the correct delay δ_{ik}: in other words, a peak of this function denotes a high coherence between the two given signals, at the corresponding lag.

A specific work on source localization in a noisy and reverberant environment has shown that the CSP analysis is the most accurate TDE technique, when compared with two other techniques, namely the Normalized Crosscorrelation and Adaptive LMS filtering [5].

Once this function is computed for each microphone pair, for each frame, and for each lag, a very simple peak-picking algorithm can be applied to derive the interchannel delay estimate, as described in [6]. Given an array geometry consisting of two microphone pairs, as in our case, the two delay estimates result by the application of the CM-based TDE technique to each pair. Then, source position is located as the intersection between direction estimates.

III. ENHANCEMENT

3.1 Delay-and-add beamformer

Once the relative delay δ'_{0k} of wavefront arrival between microphone 0 and k has been estimated, a simple technique to reconstruct an enhanced version $r'(t)$ of the acoustic message is based on the following Time Delay Compensation (TDC) relationship:

$$r'(t) = \sum_{k=0}^{M-1} \alpha'_k s_k(t + \delta'_{0k}) \qquad (4)$$

where α'_k is inverse proportional to the distance between the k-th microphone and the estimated source position.

Note that a direct computation of the corresponding spectrum can be accomplished by exploiting spectra previously processed in the localization procedure.

3.2 Mean Spectrum based Enhancement

Given the signal resulting by the previously described processing , let us indicate with $S_b(n,m)$ the discrete m-th component of its power spectrum evaluated for the n-th frame. A further enhanced version of this spectrum can be obtained as follows:

$$S_e(n,m) = \begin{cases} D, & \text{if } D \geq \beta S_b(n,m); \\ \beta S_b(n,m), & \text{otherwise.} \end{cases}$$

where: $D = [S_b(n,m) - \kappa \widehat{S_b}(m)]$, κ defines the overestimation factor of the average spectrum components, β represents the spectral flooring, and $\widehat{S_b}(m)$ is the average m-th spectrum component, evaluated on the whole utterance.

It is worth noting that the resulting enhanced spectrum represents a distorted version of the original one. However, suitable choices of the parameters β, κ can lead to an effective preprocessing for the recognizer, as discussed in the following.

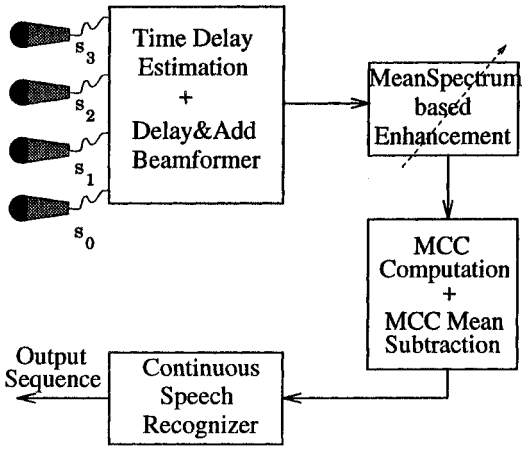

Figure 2: *A block diagram representation of the microphone array-based recognition system.*

Very similar MSE procedures were described in [9, 10]: in those cases, the enhancement processing was applied to the mel-scale filter bank output, using a running average of the generic spectrum component; in one case, β, κ coefficients were frequency dependent. These alternatives will be investigated in the future.

3.3 Cepstrum Mean Subtraction

A common simple approach to reduce the mismatch between training and testing conditions, when dealing with noisy input speech, is given by cepstral mean normalization. Here, Cepstrum Mean Subtraction (CMS) consists in computing the sample mean of the mel-cepstrum vector over the utterance; then, the mean vector is subtracted from every cepstrum vector, without distinguishing noise from speech. Recent experiments [3, 11, 12] have shown the effectiveness of this method.

Even if a spectrum normalization is accomplished by the MSE procedure described in the previous section, the combined use of these techniques introduced further benefits, as shown following.

IV. RECOGNITION SYSTEM

4.1 Acoustic Processing

Given a noisy utterance acquired through the microphone array, an enhanced version was derived by the processing described in Sections 3.1 and 3.2. During recognition, either the enhanced signal or its original replica (i.e. a unique microphone input) were preemphasized by using a digital filter having transfer function $H(z) = 1 - 0.95 \times z^{-1}$ and then processed without any start-end point detection. The signal was blocked into frames by applying a 20 ms Hamming window every 10 ms. For each frame, 8 Mel scaled Cepstral Coefficients (MCCs) were extracted, using a 24-channel filter-bank. The log-energy was computed and normalized with respect to the maximum value in the sentence. As previously mentioned, in some experiments CMS procedure was applied to MCCs. The resulting coefficients and the normalized log-energy, together with their first and second order derivatives, computed on windows of 50 ms and 70 ms length respectively, were arranged in a single observation vector of 27 components.

4.2 HMM-based Recognition

The recognition system makes use of Continuous Density HMMs with Gaussian mixture observation densities for acoustic modeling. A set of 38 Context Independent Units (CIUs) was used with 16 mixture components per state; a baseline is described in [7, 13]. Recognizer training, based on Maximum Likelihood Estimation, was accomplished by using the segmentation and labeling available with the database APASCI. During the training phase, less used gaussians were pruned. Recognition was performed with the Viterbi algorithm on model networks, depending on the type of task (Phone Loop grammar(PL) or Word Pair grammar(WP)). For the Word Pair task, an artificial grammar of perplexity 50 was defined: here, perplexity is intended as number of successors of a given word admitted by the grammar. As shown in [13], even if the vocabulary size is approximately 1000, the resulting WP task performance is high, due to the low perplexity of the grammar. Nevertheless, it is indicative of the behavior of the enhancement scheme both in clean and in noisy conditions, as shown in next section.

V. EXPERIMENTS AND RESULTS

5.1 APASCI Corpus

The present release APASCI 2.0 includes 3900 phonetically rich utterances (pronounced by 176 speakers), automatically segmented and labelled as described in [14]. The whole corpus was divided into a training set (2140 sentences uttered by 50 males and 50 females), a development set (900 sentences uttered by 18 males and 18 females), and a test set (660 sentences uttered by 20 males and 20 females).

In this work, a small portion (39 sentences uttered by two male and two female speakers) of the test set was used. For reference purposes, the same sentences belong to the 300Core Test set described in [13]. Further, 16 sentences uttered by the same speakers, were used to tune parameters of the MSE procedure. These development and test sets include 879 and 2121 phone-like units, respectively.

5.2 Noisy Speech Corpus

Then, a database of 55 utterances was acquired in a noisy office environment: speakers and sentences were the same of the clean speech case above described. Noisy conditions were typical of an office with computers, air conditioning, etc., for a resulting SNR that was estimated between 10 dB and 15 dB. The office measured $8m$ by $8m$ by $2.8m$ and was covered with plastic

panels and there were furniture and fittings. Due to the characteristics of this room, recordings included reverberation components as well as multipath distortions; further, signals were affected by coherent noise due to secondary sources (e.g. computers).

The array consisted of two distant microphone pairs: distance between microphones of each pair was 15 cm, while distance between microphone pairs was 75 cm. The results described in the following refer to the use of utterances pronounced at one meter from the array.

5.3 System Performance

Providing system performance either in terms of localization accuracy or in terms of wavefront direction accuracy is not the main purpose of this paper: a discussion on this issue can be found in [5]. As a matter of fact, talkers at one meter distance from the array are generally located correctly with an accuracy of a few centimeters.

Noisy recognition performance are given in Table 1, for the application of different preprocessing modules to either the four microphone array signals or the single microphone signal s_0. As reference result, when clean utterances of the APASCI corpus were used as input to the recognizer a 75% Phone Accuracy(PA), and a 97.8% Word Accuracy(WA) were obtained for the PL and WP tasks, respectively.

Clearly, when the signal of a single array microphone was used as input without any preprocessing, performance fell drastically to 28.4% PA and 4.4% WA; actually, output phone sequences look like they were generated almost randomly. Other results of Table 1 show that the introduction of MSE and CMS procedures, alone or together, provided an improvement. Contribution of the TDC module deserves to be underlined: only using this processing (that is the merging of four microphone signals) leads to a Phone Accuracy higher than the combination of the other two techniques, when applied to a single microphone signal. A significant result was also obtained when combining MSE and CMS with TDC: in this case, 43% PA and 78.2% WA were obtained.

It is worth noting that preliminary experiments showed benefits in using either all this processing or each module alone, even with clean speech as input.

INPUT	U.A.%	P.C.%	W.A.%
S_0	28.4	30.5	4.4
S_0+MSE	31.9	34.9	27.3
S_0+CMS	31.4	33.4	17.2
S_0+MSE+CMS	34.7	37.3	41.7
TDC	35.8	38.2	31.6
TDC+MSE	38.9	41.4	49.1
TDC+CMS	39.5	43.2	50.1
TDC+MSE+CMS	43.0	46.0	78.2

Table 1: *Recognition performance on the noisy speech test set, using different enhancement combinations.*

VI. FUTURE WORK

Even if a considerable improvement was obtained by means of simple enhancement techniques, performance of a speaker-independent recognizer in a generic noisy and reverberant environment is not yet satisfactory.

In future work we expect to improve talker localization by using more microphones and different array geometries, allowing to obtain a better beamforming. We plan to collect a larger database of noisy speech and to investigate performance in different (also more severe) environmental conditions, using the described enhancement processing. Next, other cepstral-based compensation or recognizer adaptation techniques will be investigated [9, 11, 12].

References

[1] S. Furui, "Toward Robust Speech Recognition under Adverse Conditions", In *Proceeding ESCA Workshop on Speech Processing in Adverse Conditions*, November 1992, pp. 31-42.

[2] B. H. Juang, "Speech Recognition in Adverse Environments", *Computer Speech and Language* (1991) 5, pp. 275-294.

[3] C. Che, Q. Lin, J. Pearson, B. de Vries, and J. Flanagan, "Microphone Arrays and Neural Networks for Robust Speech Recognition", *ARPA Workshop on Human language Technology*, NJ, March 1994, pp. 321-326.

[4] H. F. Silverman, S. E. Kirtman, "A Two-stage Algorithm for Determining Talker Location from Linear Microphone Array Data", *Computer Speech and Language* (1992) 6, pp. 129-152.

[5] M. Omologo, P. Svaizer "Acoustic Event Localization using a Crosspower-Spectrum Phase based Technique", *Proc. ICASSP*, Adelaide 1994, pp. II273-II276.

[6] M. Omologo, P. Svaizer, "Talker Localization and Speech Enhancement in a Noisy Environment using a Microphone Array based Acquisition System", *Proceedings Eurospeech*, Berlin, September 1993, pp. 605-609.

[7] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, M. Omologo, "Speaker Independent Continuous Speech Recognition using an Acoustic-Phonetic Italian Corpus", elsewhere in these proceedings.

[8] C. H. Knapp, G. C. Carter, "The Generalized Correlation Method for Estimation of Time Delay", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. ASSP-24, n. 4, August 1976.

[9] J.A. Nolazco Flores, S.J Young, "Continuous Speech Recognition in Noise using Spectral Subtraction and HMM Adaptation", *Proc. ICASSP*, Adelaide 1994, pp. I409-I412.

[10] D. Van Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System", *Computer Speech and Language* (1989) 3, pp. 151-167.

[11] F. Liu, P. Moreno, R.Stern, A. Acero, "Signal Processing for Robust Speech Recognition", *ARPA Workshop on Human language Technology*, Plainsboro NJ, March 1994, pp. 309-314.

[12] A. Anastasakos, F. Kubala, J. Makhoul, R. Schwartz, "Adaptation to new Microphones using Tied-Mixture Normalization", *ARPA Workshop on Human language Technology*, Plainsboro NJ, March 1994, pp. 304-308.

[13] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo. "A Baseline of a Speaker Independent Continuous Speech Recognizer of Italian", In *Proceedings of Eurospeech-93*, Berlin, September 1993, pp. 847–850.

[14] F. Brugnara, D. Falavigna, and M. Omologo. "Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models". *Speech Communication*, Vol. 12, no. 4, (1993),pp. 357–370.