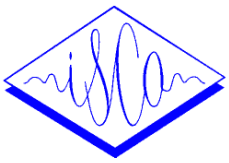

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

ROBUST CONTINUOUS SPEECH RECOGNITION USING A MICROPHONE ARRAY

D. Giuliani, M. Matassoni, M. Omologo and P. Svaizer

e-mail: {giuliani,matasso,omologo,svaizer}@itc.irst.it
IRST-Istituto per la Ricerca Scientifica e Tecnologica
I-38050 Povo (Trento), ITALY

ABSTRACT

This paper focuses on the use of microphone arrays for speaker independent continuous speech recognition in real environment. An array of four omnidirectional microphones was placed at 1.5 m distance from the talker; given the array signals, a Time Delay Compensation (TDC) technique was applied to provide a beamformed signal, that is shown effective as input to a Hidden Markov Model (HMM) based recognizer. The paper also refers to three enhancement/normalization techniques that are employed to obtain a better alignment between the new noisy conditions and the clean ones, under which the training was carried out. In particular, a phone HMM adaptation technique seems to be the most promising for developing a scenario of rapid speaker/environment/channel adaptation of a real-time hands-free recognizer.

1. INTRODUCTION

When moving the speech recognition technology from laboratory to "real world" we have to deal with a mismatch between training and testing acoustic conditions. Without any adaptation the system performance can fall down drastically. One of the main reasons is the environmental noise that corrupts the speech signal. Additional sources of mismatch are the interaction attitude adopted by the speaker, that can vary considerably from laboratory to "real world", and the alterations in the acquisition channel (such as different microphone, telephone line, etc.).

Retraining speech recognizers for every new condition is a time consuming procedure and would not solve all these problems. In this paper we refer to a target scenario that includes (1) an acquisition system, based on a microphone array, able to locate talker [1] and to reduce influence of undesired environmental components, and (2) a speaker independent continuous speech recognizer, trained on clean speech and self-adapting to new noisy conditions in real time.

Two enhancement techniques are employed, one exploiting the microphone array based acquisition system to derive a beamformed signal through a Time Delay Compensation (TDC) technique, the other that emphasizes the most prominent spectrum components of the resulting signal. Further, two compensation techniques are explored: one is called Phone Dependent Linear Regression (PDLR) and realizes a phone-dependent acoustic feature mapping; the other (HMM Adaptation) transforms phone HMMs, previously trained on clean speech, given a new small compensation database. A block diagram of the overall system is given in Figure 1.

In a previous work, some preliminary results were given [2], that showed advantages attained using the array based beamformed signal as input to the speech recognizer. As described in the following, other experiments have been carried out on a new multichannel continuous speech corpus, collected in a noisy and reverberant office; new results confirms that trend, whatever enhancement/normalization technique is employed, but in particular using an HMM adaptation.

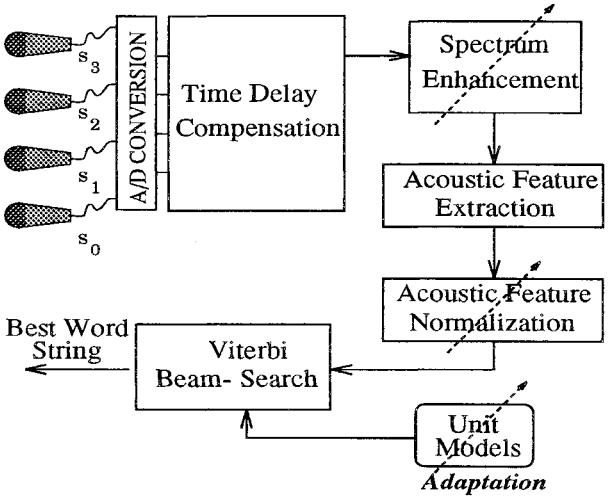

Figure 1: A block diagram representation of the microphone array-based recognition system.

2. SPECTRUM ENHANCEMENT

2.1. Time Delay Compensation

The use of a microphone array for hands-free speech recognition relies on the possibility of obtaining a signal of improved quality, compared to the one recorded by a single microphone.

Let us assume that a talker generates an acoustic event $s(t)$ that is acquired by microphones $0, \ldots, (M-1)$ as signals $s_0(t), \ldots, s_{M-1}(t)$. Signals sampled by the acoustic sensors i and k are characterized by the relative delay δ_{ik} of the direct wavefront arrival. Time delay estimation is a critical issue in noisy and reverberant conditions: in this work we adopted a CrosspowerSpectrum Phase (CSP) technique, that has been shown to be effective for acoustic event detection and location [1].

Once each relative delay $\hat{\delta}_{0k}$ of direct wavefront arrival between microphone 0 and k has been estimated, the simplest technique to reconstruct an enhanced version $\hat{s}(t)$ of the acoustic message is based

on a Time Delay Compensation (delay and sum beamformer):

$$\hat{s}(t) = \frac{1}{M} \sum_{k=0}^{M-1} s_k(t + \hat{\delta}_{0k}). \quad (1)$$

The frequency domain counterpart of this operation can be easily performed exploiting the spectra already derived in the CSP processing.

Using few array microphones, linearly placed, enhancement capabilities can be limited by the presence of noise sources in the steered direction (as well as by the arrival of reverberation from the same direction). At this moment, we are exploring the use of different geometries, more microphones, and more selective reconstruction techniques [3], to further improve recognition performance.

2.2. Discriminative Spectrum Weighting Function

A suitable piecewise linear weighting of the spectrum components is effective for spectrum enhancement. Using this technique, each frame of the input utterance is classified either as "noisy" (frame non containing speech signal) or as "speech ". Then, a simple attenuation or a non-linear weighting function is applied to spectrum components of "noisy" or "speech" frames, respectively. A more detailed description of this technique is given in [2]: the resulting acoustic features seem to be more robust to background noise, but performance improvement is generally limited. Following we will refer to this technique as *DSWF*.

3. COMPENSATION METHODS

3.1. Phone Dependent Linear Regression

Recently a piecewise-linear mapping was introduced for channel normalization [4, 5] in order to exploit the non-linear relationship between acoustic spaces.

In this work, a phone dependent linear mapping is employed. Given the simultaneous recordings available for each speech stimulus, two sequences of corresponding feature vectors are derived that represent the clean (i.e. close-talk) and the noisy (i.e. array microphone) feature spaces. Applying to each signal an automatic phone segmentation and labeling procedure [6] (assuming that the text uttered is known), a set of noisy and clean feature vector pairs is individuated for each phone: these vectors are used to estimate the phone dependent linear mapping.

During recognition, channel normalization is performed by mapping noisy features into the clean feature space. Just MCCs (Mel Cepstral Coefficients) and energy are transformed, while first and second order derivatives are computed from transformed parameters. For each frame, the optimal transformation is obtained by a suitable weighted linear combination of the estimated phone dependent linear transformations as follows:

$$\tilde{x} = \sum_{c=0}^{C-1} p(c|x)[\mu_{\tilde{x}}^{(c)} + A^{(c)^T} \cdot (x - \mu_x^{(c)})] \quad (2)$$

where: x is the current input vector; $\tilde{x}$ is the corresponding transformed vector; $\mu_x^{(c)}$ and $\mu_{\tilde{x}}^{(c)}$ are the mean vectors for the phone c in the clean and in the noisy feature spaces, respectively; $A^{(c)}$ is the phone dependent linear transformation matrix for the phone c, and C is the number of phones. Finally, $p(c|x)$ is the posterior probability that x corresponds to a realization of the phone c: this probability is computed estimating a likelihood $p(x|c)$, using a Gaussian probability density function for each phone, and then applying the following relationship:

$$p(c|x) = p(x|c)p(c) / \sum_{c=0}^{C-1} p(x|c)p(c) \quad (3)$$

where $p(c)$ indicates the prior probability of the phone c.

Following we will refer to this technique as *PDLR*.

3.2. HMM Adaptation

This technique is largely used for speaker adaptation purposes [7]. Basically it consists in reestimating Gaussian means of a pre-trained set of HMMs using a small amount of data. In this way, speaker and channel adaptations are performed at the same time (since both of them change from training to test conditions).

In this work, reestimation of Gaussian means was realized according to the Maximum Likelihood (ML) criterion and the Viterbi algorithm. No interpolation between old and new models was applied.

Following we will refer to this technique as *Ada*.

4. RECOGNITION SYSTEM

4.1. Acoustic Processing

Each signal is preemphasized and blocked into frames of 10 ms duration. For each frame, 8 Mel scaled Cepstral Coefficients (MCCs) are extracted and normalized by subtracting the MCC means computed on the whole utterance. The log-energy is also computed and normalized with respect to the maximum value in the sentence. The resulting coefficients and the normalized log-energy, together with their first and second order derivatives, are arranged into a single observation vector of 27 components. During test, this acoustic processing is applied either to the single microphone signal or to the output of the TDC module.

4.2. HMM-based Recognition

A set of 33 Context Independent Units (CIUs) are modeled by means of continuous density HMMs. A left-to-right topology with three states (without skip among states) is adopted for all the CIUs with the exception of the "silence" unit, for which a single state topology is used. Output distribution probabilities are modeled by means of mixtures having 16 Gaussian components with diagonal covariance matrix.

Recognizer training, based on maximum likelihood estimation, is accomplished by using the segmentation and labeling available with the database APASCI described below. During the training phase, less used Gaussians are pruned. Recognition is performed with the Viterbi algorithm on finite state networks, depending on the type of task.

5. SPEECH DATABASES

5.1. APASCI Corpus

The APASCI corpus was collected for the development of a speaker independent continuous speech

	CTMic	*MicArray*	*Ch0Mic*
Baseline	71.0	40.1	32.3
DSWF	70.3	45.3	39.2
PDLR	-	52.3	45.9
Ada	79.0	60.4	52.4
PDLR + Ada	-	60.5	53.3
All	-	60.2	53.5

Table 1: Phone Accuracy for the PL task, measured on 240 test sentences, using different microphones for acquisition and different solutions for enhancement/normalization. *All* indicates the joint use of all the given techniques.

	CTMic	*MicArray*	*Ch0Mic*
Baseline	81.0	50.3	31.5
DSWF	81.2	58.3	43.1
PDLR	-	62.9	50.4
Ada	85.5	73.8	65.8
PDLR + Ada	-	71.6	62.7
All	-	70.4	61.7

Table 2: Word Accuracy for the WL task, measured on 240 test sentences, using different microphones for acquisition and different solutions for enhancement/normalization. *All* indicates the joint use of all the given techniques.

recognizer for italian language whose baseline is described in [8]. The present release includes a training set that consists of 2140 sentences (uttered by 50 males and 50 females). The speech material was acquired in a quiet room, using a Sennheiser close-talk cardioid microphone.

5.2. Multichannel Speech Corpus

In order to measure performance using noisy speech material, acquired either with a close-talk microphone or with a distant microphone array, a new corpus was collected in an office environment. Due to the characteristics of the room, recordings included reverberation components, as well as coherent noise due to secondary sources (e.g. computers, air conditioning, etc).

Multichannel recording of each utterance was accomplished by using a close-talk cardioid microphone (following called *CTMic*) and a linear microphone array (following called *MicArray*) situated in front of the speaker at 150 cm distance. The array consisted of four microphones: distance between microphones was 30 cm. For comparison purposes, a single microphone (namely *Ch0Mic*) of the array will be considered as an independent acquisition channel.

Acquisitions were carried out synchronously for all the input channels at 16kHz sampling frequency, with 16 bit accuracy. Signal to Noise Ratio (SNR) referred to *CTMic* and to *Ch0Mic* material was 25 dB and 10 dB, respectively. Each value was measured as ratio between speech energy and noise energy, according to the speech-noise classification provided by the automatic segmentation and labeling system.

Eighty sentences were uttered by four speakers (2 males and 2 females). For each speaker, a development set and a test set were defined, that consist in 20 sentences and 60 sentences, respectively. The resulting test set includes 789 words (13492 phone-like units) and is characterized by a word dictionary size of 343. The development sets were employed to tune parameters of the *DSWF*, *PDLR*, and *Ada* modules.

As mentioned above, *CTMic* material was automatically segmented and labeled: resulting segmentation was exploited to document the corresponding *MicArray* material as well.

6. EXPERIMENTS AND RESULTS

6.1. System Performance

Given the multichannel speech material described above, a set of experiments was carried out, whose performance is reported in terms of Phone Accuracy (PA) and Word Accuracy (WA). Phone Accuracy was evaluated using a Phone Loop (PL) grammar without any phone statistics or phonotactic constraint. Word accuracy was measured given a Word Loop (WL) grammar: this grammar has a single state and a self-loop per word; the resulting perplexity is 343 (i.e. the size of the dictionary).

In Table 1, PAs are reported for the different enhancement and compensation techniques, used either separately or jointly. Corresponding WAs for the WL task are reported in Table 2. For both tables, results were derived averaging performance obtained on each of the four speakers.

All the experiments emphasized benefits due to the use of the microphone array (and of the related beamforming processing) with respect to the use of a single microphone of the same array. Note that the main differences we observed, comparing a beamformed signal with the corresponding *Ch0Mic* one, were: (1) a more coherent redistribution of the spectrum components (both in the power spectrum and in the phase spectrum domain) related to the formant bands, and (2) a perceivable low-pass effect. As a result, the new signal seems less reverberant and less noisy, even if the corresponding SNR does not change substantially from that evaluated on the single microphone signal.

Then, one can note that applying the *PDLR* technique (52.3% PA, 62.9% WA) to the beamformed signal improves performance obtained using the *DSWF* technique (45.3% PA, 58.3% WA), but results are not comparable to those provided using the *Ada* technique (60.4% PA, 73.8% WA). Actually, the *DSWF* performs a non-linear transformation to emphasize higher-energy spectral bands (and, at the same time, to deemphasize low-energy ones), no matter which is the phonetic content. On the other hand, the *PDLR* technique realizes a phone dependent feature mapping that should represent a more powerful transformation than the previous one. Nevertheless, both of them are not so effective as a Phone model adaptation.

Further, at this moment the joint use of these techniques does not show advantages with respect to the use of the *Ada* technique alone. A better combination could be accomplished by performing an optimal global tuning as well as a re-training of clean-speech based models by using the same acoustic processing. This issue will be investigated next.

	CTMic	*MicArray*
NoAdaptation	71.0	40.1
LeaveOneOut	71.3	51.9
Multi – Speaker	75.8	56.2
Single – Speaker	79.0	60.4

Table 3: Phone Accuracy using the *Ada* technique with different use of compensation material.

	CTMic	*MicArray*
NoAdaptation	81.0	50.3
LeaveOneOut	78.7	65.1
Multi – Speaker	82.8	71.6
Single – Speaker	85.5	73.8

Table 4: Word Accuracy using the *Ada* technique with different use of compensation material.

6.2. Influence of the adaptation material

Experiments showed that the joint use of the microphone array and the *Ada* normalization technique allows to obtain a 73.8% WA (60.4 % PA) to be compared with 31.5%WA (32.3% PA) obtained using a single microphone of the array itself. The reference system, based on the use of a close-talk microphone, provided a corresponding 85.5% WA (79.0% PA).

However, this adaptation involves a realignment concerning both the channel acquisition and the new speaker. This section refers to a deeper analysis in order to distinguish a speaker dependent improvement component from a channel/environment dependent one, given the overall improvement.

Table 3 and Table 4 report performance comparisons when the *Ada* technique is applied to different compensation material obtained as follows: (1) *Single – Speaker* mode corresponds to use 20 adaptation sentences pronounced by a speaker before a test on his own test material (then performance is averaged on all the speakers); (2) *Multi – Speaker* mode corresponds to use 80 adaptation sentences (20 utterances by each speaker) before performing recognition on the whole test set; (3) *LeaveOneOut* mode corresponds to use, for a given speaker, 60 adaptation sentences (20 utterances by each of the remaining speakers).

The *LeaveOneOut* experiments show that the improvement obtained using the *CTMic* microphone, mainly depends on the adaptation to the new speaker; on the other hand, using the microphone array as input, a consistent improvement is due also to the phone model adaptation to the new channel/environment conditions. Finally, note that the *Multi – Speaker* performances are close to the *Single – Speaker* ones, even if this result deserves further investigation using more speakers and different size of compensation material.

7. FUTURE WORK

Results given above show that a simple combination of a microphone array based processing and a HMM based recognizer, trained under clean conditions and rapidly adapted to new conditions, can provide encouraging results in a real environment task. We expect significant improvement further working on each module of the overall system and on global tuning of the employed techniques.

Even if multichannel data collection is a time-consuming activity, it represents a fundamental way to evaluate system performance in real environment. In order to better assess the results given above, a new multichannel database is being planned. A large number of utterances will be collected with different talker positions, under different noisy conditions, and employing different array geometries.

Finally, we plan to investigate the use of an alternative normalization module, based on a multilayer perceptron, able to transform acoustic features obtained from distant-talk microphone array to those corresponding to close-talk microphone, as suggested by [9].

ACKNOWLEDGEMENT

The authors would like to thank Bianca Angelini for her help in collecting the multichannel corpus.

References

[1] M. Omologo, P. Svaizer, "Acoustic Event Localization using a Crosspower-Spectrum Phase based Technique", *Proc. ICASSP*, Adelaide 1994, Vol. 2, pp. 273-276.

[2] D. Giuliani, M. Matassoni, M. Omologo, P. Svaizer, "Hands Free Continuous Speech Recognition in Noisy Environment using a Four Microphone Array", *Proc. ICASSP*, Detroit, April 1995, Vol. 2, pp. 860-863.

[3] E.E. Jan, P. Svaizer, J.L. Flanagan, "A Database for Microphone Array Experimentation", elsewhere in these proceedings.

[4] L. Neumeyer, M. Weintraub, "Probabilistic Optimum Filtering for Robust Speech Recognition", *Proc. ICASSP*, Adelaide, Australia, 1994, Vol. 1, pp. 417–420.

[5] L. Neumeyer, M. Weintraub, "Robust Speech Recognition in Noise using Adaptation and Mapping Techniques", *Proc.ICASSP*, Detroit, April 1995, Vol. 1, pp. 141 - 144.

[6] 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.

[7] D. B. Paul, "The Lincoln Large-Vocabulary Stack-Decoder Based HMM CSR", *Proc. ARPA* Workshop, Plainsboro, NJ, March, 1994.

[8] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, M. Omologo, "Speaker Independent Continuous Speech Recognition using an Acoustic-Phonetic Italian Corpus", *Proc. ICSLP*, Yokohama, September 1994, Vol. 3, pp. 1391–1394.

[9] 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.