HANDS FREE CONTINUOUS SPEECH RECOGNITION IN NOISY ENVIRONMENT USING A FOUR MICROPHONE ARRAY

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

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

ABSTRACT

This paper describes recent advances on the use of HMM based technology, for speaker independent continuous speech recognition, in noisy environment, under hands free interaction mode.

For this purpose an array of four ominidirectional microphones is employed as acquisition system. The processing of phase information in the Cross-power Spectrum provides the capability both of locating talker position and of reconstructing an enhanced speech spectrum.

Here, two enhancement techniques are described, that allow recognition improvement in the case of clean input speech as well as under different adverse conditions. Results refer to the use of a new multichannel corpus, collected in real environment by microphone array as well as close-talk microphone.

1. INTRODUCTION

The ability of speech recognition systems to deal with diversified environmental conditions, speaker interaction modes, and acquisition channels is still far to be achieved.

When there is mismatch in training and testing conditions, system performance falls down drastically. Sometimes, the main reason is environmental noise, that is present in the speech signal; a related one is the concurrent interaction mode adopted by the speaker, that can vary considerably from laboratory to "real world" [1].

Retraining speech recognizers for every new condition is a time consuming procedure and would not solve the problem. In this work, we jointly consider two aspects of the problem: noise adaptation and talker-focused acquisition. A target scenario is devised including: a) an acquisition system based on a microphone array able to locate talker and to reduce influence of undesired environmental components, b) a speaker independent continuous speech recognizer trained on clean speech and self-adapting in real-time to new noisy conditions.

The present version of the detection and location system is described in [2], while an introduction to the recognition system and preliminary results about its use in adverse conditions can be found in [3]. A block diagram of the overall system is given in Figure 1.

Time Delay Compensation (TDC) module performs talker location and source beamforming using only four microphones. Improved performance is expected by using a more sophisticated acquisition system with a larger number of sensors in order to achieve better spatial selectivity in sound pick-up.

In the remainder of the paper, two spectrum enhancement techniques are described, that derive from those introduced in [3]. Recognition results refer to experiments on a new multichannel corpus, collected in an office under different noisy conditions, using both the array and a close-talk microphone.

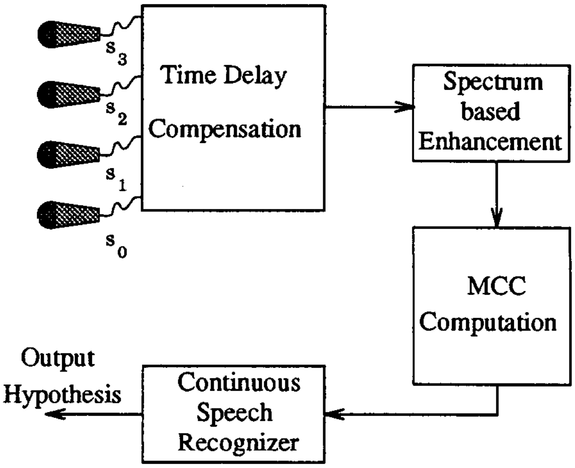

Figure 1: *A block diagram representation of the microphone array-based recognition system.*

2. TIME DELAY COMPENSATION

The use of a microphone array for speech recognition relies on the possibility of obtaining a signal of improved quality, compared to the signal acquired by a single microphone, when operating in adverse conditions such as in noisy and reverberant environment and with a distant talker.

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 acquired 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 [4] that has been shown to be effective for acoustic event detection and location [2].

0-7803-2431-5/95 $4.00 © 1995 IEEE

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):

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

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

3. SPECTRUM ENHANCEMENT

3.1. Spectrum Weighting Function

A suitable piecewise linear weighting of the spectrum components is effective for spectrum enhancement [3, 5, 6].

Let us indicate $S(n,m)$ as the discrete m-th component of the power spectrum at n-th time-frame, and $\overline{S}(m)$ its average evaluated on the whole utterance. We introduce the following new non-linearly weighted spectrum:

$$S_e(n,m) = \frac{\gamma\alpha\overline{S}(m)}{\frac{(\gamma-\beta)}{\gamma\alpha\overline{S}(m)} S(n,m) + 1} + \gamma(S(n,m) - \alpha\overline{S}(m)) \tag{2}$$

where α, β, and γ are parameters fixed on empirical basis. An example of the corresponding weighting function is reported in Figure 2. Following we will refer to this technique as SWF (Spectrum Weighting Function).

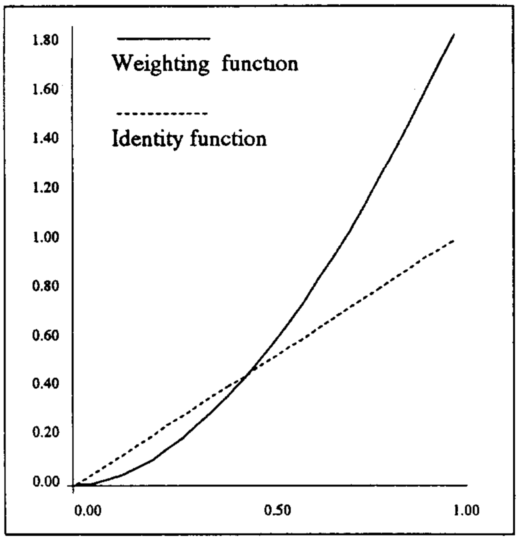

Figure 2: *Example of weighting function, for input values between 0.0 and 1.0.*

3.2. Discriminative Spectrum Weighting Function

Another version of the enhancement technique proposed in the previous section is described following.

If n-th frame is classified as "noise" (frame non containing speech signal) then: $S_e(n,m) = \delta S(n,m)$, with $0 < \delta < 1$. Otherwise, (2) is applied by replacing $\overline{S}(m)$ with $[\overline{S}_{ns}(m) + c\sigma_{ns}(m)]$, where $\overline{S}_{ns}(m)$ and $\sigma_{ns}(m)$ are the average and standard deviation of $S(n,m)$ in "noise" frames of the utterance. Initially, $\overline{S}_{ns}(m)$ and $\sigma_{ns}(m)$ are evaluated on first l frames, which are assumed to be always "noise" (in this work, l was fixed to 20). Then, the estimation proceeds until a "speech" frame is encountered.

In order to distinguish among "speech" and "noise" frames, a simple heuristic rule is adopted. Given a frame, the number of spectrum components is evaluated, for which $S(n,m) > [\overline{S}_{ns}(m) + c\sigma_{ns}(m)]$: if this number is bigger than a given threshold Th, the frame is classified as "speech" otherwise as "noise". In this work, c was fixed to 3, while Th was estimated from the first l frames of the utterance.

Following we will refer to this technique as DSWF (Discriminative Spectrum Weighting Function).

4. RECOGNITION SYSTEM

4.1. Acoustic Processing

Each signal is 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 is blocked into frames by applying a 20 ms Hamming window every 10 ms. For each frame, 8 Mel scaled Cepstral Coefficients (MCCs) are extracted, using a 24-channel filter-bank. MCCs are normalized by subtracting the MCC means computed on the whole utterance, useful for compensation of channel transfer function effects [3]. 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, computed on windows of 50 ms and 70 ms length respectively, are arranged in 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 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 an ergodic topology with a single state 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 present release of the italian corpus APASCI 2.0 [7] includes 3900 phonetically rich utterances (pronounced by 176 speakers), automatically segmented and labeled as described in [8]. 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). The speech material was acquired in a quiet room.

This corpus was collected for the development of a speaker independent continuous speech recognizer for italian language whose baseline is described in [7].

5.2. Multichannel Speech Corpus

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

	CloseTalkMic	*Ch0Mic*
Clean	30.1	11.9
Noise1	26.3	6.5
Noise2	21.8	0.2

Table 1: Mean SNR in dB using acquisition channels *CloseTalkMic* and *Ch0Mic* with three different noisy conditions (namely *Clean*, *Noise1* and *Noise2*).

A multichannel recording of each utterance was provided by a close-talk cardioid AKG D3700 microphone (following called *CloseTalkMic*) and a linear microphone array (following called *MicArray*) situated in front of the speaker at 150 cm distance. 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. Acquisitions were carried out, synchronously for all the input channels at 24kHz sampling frequency, with 16 bit accuracy. Downsampling to 16kHz was then applied.

Ten sentences were uttered by each of four speakers (2 males and 2 females) in three different noisy conditions (sentences were different speaker by speaker). For each condition, the resulting test set consists of 371 words (2189 phone-like units). The first acquisition session was performed under normal office conditions, the second and the third ones were accomplished introducing two levels of background noise in the office (two loudspeakers reproduced recordings of real environment noise). Following, the mentioned conditions will be called *Clean*, *Noise1* and *Noise2*. For comparison purposes, a single microphone (namely *Ch0Mic*) of the array will be also considered as an independent acquisition channel. *CloseTalkMic* speech material was automatically labeled and segmented [8], in a coherent way with the APASCI notation, and used to document *MicArray* material as well.

Table 1 reports mean Signal to Noise Ratios (SNRs) for different acquisition channels and noisy conditions. SNR was measured as ratio between speech energy and noise energy, according to the speech-noise classification provided by segmentation and labeling.

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 Recognition Rate (WRR). Phone Accuracy was evaluated using an unit loop grammar without any phone statistics or phonotactic constraints. Word Recognition Rate was measured for two different tasks: Word Pair grammar (WP) and a Word Loop grammar (WL). For the WP task, an artificial grammar having perplexity 50 was built (perplexity is intended as number of successors of a given word in the grammar). The WL grammar has a single state and a self-loop per word. In this case, the perplexity is 371 (i.e. the size of the dictionary).

In Table 2, PA is reported for each acquisition channel and noisy condition. Results were obtained using both the baseline system and the above mentioned enhancement techniques.

Experimental results, in terms of WRR for the WP task and the WL task, are reported in Table 3 and 4, respectively. WRR are given only for the *Clean* test sets.

6.2. Discussion

Results of Table 2 suggest a first comment on the use of close-talk microphones. Under the most adverse condition *Noise2*, performance falls down of 8% PA even using the best enhancement method. This result can represent a reference for the other experiments, where distant-talk microphone were employed, and can be probably related mainly to stressed speech phenomena. In fact, talkers uttered at a very small distance from the close-talk microphone, but were influenced by the environmental noise, causing effects like Lombard speech not considered in the enhancement technique development. This subject deserves further studies.

Still looking at Table 2, results show performance improvement, due to the use of spectrum enhancement and microphone array processing, that leads from 27.9% to 42.7% PA in the case of *Clean* test set. The corresponding improvement from 19.1% to 81.4% WRR (see Table 3) and from 15.9% to 49.8% WRR (see Table 4), for the WP task and the WL task, respectively, confirm the joint benefit of these two methods.

Some small discrepancies in results attained using SWF and DSWF enhancement techniques can be caused by a non-optimal choice of parameters (described in Section 3.1), that were not tuned to the specific acquisition conditions of each test set.

	CloseTalkMic			*MicArray*			*Ch0Mic*		
	Clean	*Noise1*	*Noise2*	*Clean*	*Noise1*	*Noise2*	*Clean*	*Noise1*	*Noise2*
Baseline	71.9	67.6	58.1	33.2	27.4	22.8	27.9	24.0	21.2
SWF	69.4	68.7	62.0	37.1	27.7	21.2	30.1	24.1	18.8
DSWF	72.0	69.2	64.1	42.7	35.4	28.1	36.3	31.9	24.7

Table 2: Phone recognition rate in terms of Phone Accuracy for the baseline system and using the enhancement methods (namely *SWF* and *DSWF*). For each acquisition channel (namely *CloseTalkMic*, *MicArray* and *Ch0Mic*) three noisy conditions are considered (*Clean*, *Noise1* and *Noise2*).

	CloseTalkMic	*MicArray*	*Ch0Mic*
Baseline	97.5	54.9	19.1
SWF	97.5	80.6	50.1
DSWF	98.1	81.4	57.4

Table 3: Word Recognition Rate for the Word Pair task. Using the baseline system alone and using the enhancement methods. WRR refers to *Clean* test for each acquisition channel (namely *CloseTalkMic*, *MicArray* and *Ch0Mic*).

	CloseTalkMic	*MicArray*	*Ch0Mic*
Baseline	75.7	27.2	15.9
SWF	76.8	47.1	32.8
DSWF	74.9	49.8	34.7

Table 4: Word Recognition Rate for the Word Loop task. Using the baseline system alone and using the enhancement methods. WRR refers to *Clean* test for each acquisition channel (namely *CloseTalkMic*, *MicArray* and *Ch0Mic*).

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, can provide encouraging results in a real environment task. We expect significant improvement further working on each module of the overall system. In particular, multichannel data collection, unit model adaptation, and acoustic feature extraction seem to be issues to address immediately.

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.

A second activity, under way, refers to the use of unit model adaptation. For this purpose, we intend to exploit segmentation and labeling of close-talk microphone material to adapt each unit model, separately and by fast techniques. Some promising results have been recently obtained with this approach.

Finally, we are investigating the use of a normalization module, based on Artificial Neural Network, able to transform acoustic features obtained from distant-talk microphone array to those corresponding to close-talk microphone, as suggested by [9].

REFERENCES

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

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

[3] D. Giuliani, M. Omologo, P. Svaizer, "Talker Localization and Speech Recognition Using a Microphone Array and a Cross-PowerSpectrum Phase Analysis", *Proc. ICSLP*, Yokohama, September 1994, Vol. 3, pp. 1243–1246.

[4] 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, No. 4, August 1976.

[5] D. Van Compernolle, "Noise Adaptation in a Hidden Markov Model Speech Recognition System", *Computer Speech and Language*, No. 3, pp. 151-167, 1989.

[6] J.A. Nolazco Flores, S.J. Young, "Continuous Speech Recognition in Noise using Spectral Subtraction and HMM Adaptation", *Proc. ICASSP*, Adelaide, April 1994, Vol. 1, pp. 409-412.

[7] 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.

[8] 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.

[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.