SPEECH ENHANCEMENT FOR SPEAKER IDENTIFICATION

Marcel Gabrea, Chakib Tadj

École de Technologie Supérieure - Electrical Engineering Department,
1100, Notre-Dame West, Montreal, Quebec, Canada H3C 1K3
{mgabrea,ctadj}@ele.etsmtl.ca

ABSTRACT

In this paper, we study the performance limits of a standard GMM speaker identification (SI) system in "adverse conditions" context using several real noises. Adaptive noise cancellation represents one such potentially effective technique and refers to a class of adaptive enhancement algorithms based on the availability of a primary input source and a secondary reference source. It will be shown that with the use of an adaptive noise cancelling called *Double Fast Recursive Least Squares*, the SI performance can approach the optimal performance system. Experiments are done on Spidre corpus corrupted by different type of noises. The performance of the SI is improved by more than 35%.

1. INTRODUCTION

It is useful to compare from an experimental point of view different automatic speaker recognizers in order to choose the one which is the most appropriate to a specific application. The best information should be obtained in carrying out experiments in real conditions for each recognizer. Unfortunately, this method is not realistic in terms of time, money, and experimental constraints. The use of speech databases aims at overcoming these difficulties, but it remains unrealistic to record one speech database for the evaluation of speech recognizers for one particular application. That is the reason why transformations of speech databases are investigated. This approach is supposed to define what transformation is to be applied to a "reference database" (Swithcboard, Spidre,...) in order to simulate new conditions. Noisy environments are representative of a large number of recognition applications, and it is necessary to have a noise database. Under the hypothesis of additive noise, noisy speech can be obtained by addition of clean speech and noise. But if one wants to be more accurate in simulating noisy speech, the Lombard effect must be taken into account. Speakers modify their speech production in presence of background noise. In this paper, we are interested in testing the limit of a GMM based speaker identification for different types of noise. The availability of a primary input source and a secondary reference source is considerated. The coupling systems is modelized as a linear time-invariant Finite Impulse Response (FIR) filters and a recursive-based adaptive filter solution to enhance the noisy speech is used. The optimum filter weight adaptation is based on a Double Fast Recursive Least Squares (DFRLS) algorithm. An acoustical comparative study with other adaptive algorithms shows the superiority of the DFRLS in terms of convergence, global and segmental SNR, informal quality and intelligibility tests [5]. In this paper it will be shown that with the use of an adaptive noise cancelling based on Double Fast Recursive Least Squares algorithm, the SI performance can approach the optimal performance system.

The rest of the paper is organized as follow: section 2 gives an overview of a GMM based SI system. Section 3 describes the techniques used for enhancement of signals degraded by additive noise. Corpus description and all the experimental framework are described in section 4. Conclusion and future experiments are addressed in section 5.

2. GAUSSIAN MIXTURE MODELS: A REVIEW

In the Gaussian Mixture Model (GMM) [6], the distribution of the parametrization speech vector of a speaker is modelized by a weighted sum of Gaussian densities:

$$p(\vec{x}\ |\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \quad \text{with} \sum_{i=1}^{M} p_i = 1 \qquad (1)$$

where $\vec{x}$ is a D-dimensional cepstral vector, λ is the speaker model, $b_i(\vec{x})$, $i = 1, \cdots, M$, are the component densities characterized by the mean $\vec{\mu}_i$ and the covariance matrix Σ_i and p_i, $i = 1, \cdots, M$, are the mixture weights. Each component density is a D-variate Gaussian Mixture function of the form:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\sum_i|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i)\} \qquad (2)$$

The model parameters $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}$ are estimated by an EM algorithm [3]. Typically, Gaussian Mixture Models (GMM) is an one state Hidden Markov Models (HMM). Having only one state instead of the standard three state phone-based HMM presents the advantage of requiring less training data.

3. ADAPTIVE NOISE CANCELLING

Several techniques exist for enhancement of signals degraded by additive noise. Adaptive noise cancellation represents one such potentially effective technique and refers to a class of adaptive enhancement algorithms based on the availability of a primary input source and a secondary reference source. The primary input source is assumed to contain speech plus additive noise.

The basic scheme of adaptive noise canceller given in [9] uses an adaptive filter based on the LMS algorithm for estimating the additive noise by filtering the reference source signal. However, the problem with this approach is that it is difficult to obtain a highly correlated noise in the primary input with the reference source signal without simultaneously obtaining a correlated speech signal. Crosstalk is consequentialy induced in the reference source signal by a closer placement of two microphones and the standard adaptive noise canceller provides verry little benefit.

The problem is the design of a structure for joint process estimation that will eliminate noise in the presence of crosstalk. Let us consider the system modeled by the diagram represented in Figure 1. The purpose is to recover the free noise speech signal $s(n)$ from the two available observations $p_1(n)$ and $p_2(n)$ in the presence of the noise signal $b(n)$.

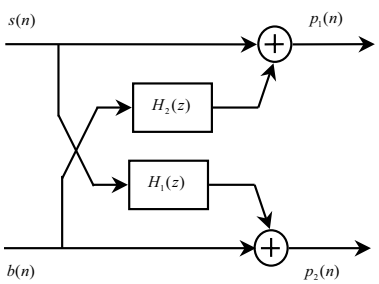

Figure 1: *Signal model.*

To recover the speech signal we use the feedback implementation of a noise canceller represented in Figure 2. We only suppose that the speech signal and the noise are statistically independents and we consider the coupling systems being FIR filters. $W_1(z)$ and $W_2(z)$ are two adaptive filters. Each one has as input the output error signal of the other filter.

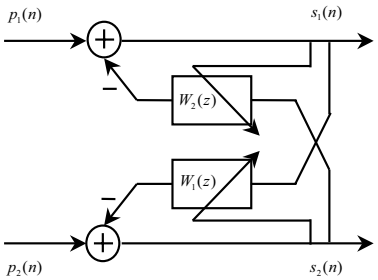

Figure 2: *Feedback implementation of the noise canceller.*

The optimum value, in the Wiener sense, of the tap-weight $W_i(z)$, $i = 1, 2$, is obtained by minimizing the mean squares errors. In the case where speech signal and the noise are statistically independents the Wiener-Hopf equations provide multiple solutions. Among all these solutions we can find the desired solution $W_i(z) = H_i(z)$, $i = 1, 2$. In this case it is easy to verify that $s_1(n) = s(n)$ and $s_2(n) = b(n)$ and it is possible to recover the signals that would have been measured at each microphone in the absence of the other source signal.

These desired solutions can be reached using a weight adaptive filters updating based on the LMS or RLS algorithm. We propose to use the Fast RLS (FRLS) algorithm for the following reasons. RLS algorithm has a rate of convergence typically an order of magnitude faster than the LMS algorithm. Among the existing versions of the FRLS algorithm which have appreciably lower computational complexity than RLS, we have opted for the numerically stabilized [2] version well adapted to non-stationary input signal like speech. On the other hand, the behaviour of these algorithms for real time applications and DSP implementations has been mastered [1]. This algorithm named Double Fast Recursive Least Squares (DFRLS) [5] can also be used for a subclass of signal separations where the direct link must be stronger than the interference link in the both channels.

4. EXPERIMENTS

4.1. Corpus Description

The Spidre corpus used consists of eighty (80) speakers; but only the 45 *target* speakers have been used. The corpus used consists of four conversation halves each from 45 claimant speakers (27 males and 18 females). The four conversations from each speaker originate from three different handsets (called *mismatch condition*) with two conversations from the same phone number (*match condition*).

4.2. Noisy and Enhanced Speech

The speech signal and the noise have been separately recorded. Four types of noise from *NOISEX* corpus are considerated: white noise, pink noise, voice babble and factory noise. The white and pink noise were acquired by sampling high-quality analog Wandel&Goltermann noise generator. The voice babble and the factory noise was acquired by recording samples from 1/2" B&K condensor microphone onto digital audio tape (DAT). The source of the babble voice is 100 people speaking in a canteen. The room radius is over two meters; therefore, individual voices are slightly audible. The sound level during the recording process was 88 dBA. The factory noise was recorded near plate-cutting and electrical welding equipment. The noises have been artificially added to the noise-free speech, so that one would master the SNR input, to obtain the noisy speech according Figure 1. The coupling systems are ten taps two FIR filters. One considers the eventual delay between the two observation signals taken into account by one of the FIR filters. The noisy signals used in SI are the signals $p_1(n)$.

The enhanced speech signals, are obtained using a recursive-based adaptive filter solution Figure 2. The optimum filter weight adaptation is based on a Double Fast Recursive Least Squares (DFRLS) algorithm. An acoustical comparative study between the Normalized LMS, other algorithms and the DFRLS algorithms shows the superiority of the noise canceller DFRLS based algorithm in terms of convergence, global and segmental SNR, informal quality and intelligibility tests [5]. Furthermore, the structure based on the coupling FIR filters permits the DFRLS algorithm to be also used as a signal separators or a signal deconvolvers rather than only a simple noise canceller. However, during high energy regions, the behaviour of the Normalized LMS is close to the DFRLS. The reason is that noise is masked by the high energy speech regions, and hence does not require complex treatment. The enhanced signals used in SI are the signals $s_1(n)$. Figure 3 shows an example of a typical signal extracted from Spidre, and the corresponding noisy and enhanced signals, in the case of a babble noise type.

4.3. Experimental Conditions

The features are a 26-dimensional vectors consisting of 12 cepstral coefficients, 12 Δ coefficients, logarithmic power and Δ logarithmic power. Analysis conditions are listed in Table 1. In all the experiments, we have used 30 seconds of speech to test the performance of the SI system.

In table 2, the minimum, maximum, mean and standard deviation SNR of noisy signals $p_1(n)$ are presented. Table 3 shows the corresponding enhanced signals $s_1(n)$, respectively.

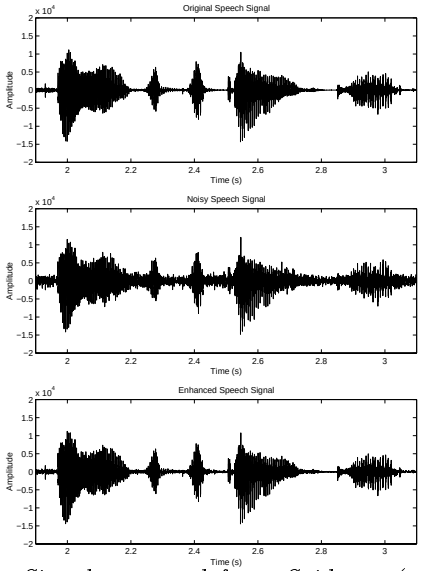

Figure 3: *Signal extracted from Spidre : (a) original signal, (b) noisy babble signal and (c) enhanced signal.*

Parameter	Value
Pre-emphasis	1-0.97 z^{-1}
Window length	25.0 ms
Window shift	10.0 ms
MFCC cepstrum order	24
Cepstral coefficient liftering	22
Cepstral mean normalization	yes
Hamming window	yes

Table 1: *Analysis conditions used for different experiments*

	Min (dB)	Max (dB)	Mean (dB)	Std (dB)
White	-6.48	15.98	5.43	4.49
Pink	-10.79	11.62	1.11	4.50
Babble	-7.32	15.22	4.61	4.51
Factory	-8.09	14.75	4.10	4.51

Table 2: Input SNR for different noises

Each speaker is modelized with a set of 35 mixtures (16 as static, 16 as dynamic and 3 as energy). Each of the mixture components has a diagonal covariance matrix. All the speech files (training and test) were run through a silence detector.

	Min (dB)	Max (dB)	Mean (dB)	Std (dB)
White	4.90	25.60	17.48	4.25
Pink	0.32	23.20	12.83	4.80
Babble	6.22	23.53	16.08	3.67
Factory	-3.10	23.19	13.34	4.88

Table 3: Output SNR for different noises

Signal Type	Performance
Original	85

Table 4: GMM Based SI performance with Spidre Corpus

4.4. Analysis and Discussion

Table 5 shows the results of the SI task on noisy and enhanced signals. By using a relative measure

$$\frac{Perf_{Enhanced} - Perf_{Noisy}}{Perf_{Optimal}} \times 100$$

we obtain an average improvement of more than 35%. All the results presented in this paper used a basic GMM implementation. Therefore no improved or re-estimated models have been used. Other improvement can be acheived by using models obtained by some transformations such as LDA, NLDA, MLLR [7, 8].

5. CONCLUSION AND FUTURE PROBLEMS

In this paper, we have presented a study of the limits of a standard GMM speaker identification system in "adverse conditions" using several real noises. We have used a Double Fast Recursive Least Squares, an adaptive noise cancelling algorithm. The results showed the efficiency of the noise canceller used by improving the SI performance by more than 35%. In our future work, we will explore other aspects of noise cancelling and various adaptation and transformation technics.

6. REFERENCES

[1] Atay, R., Baylou, P., and Najim, M., "Implementation of the Stabilized Fast Transversal Filter Algorithm on Fixed Point DSP", *ICASSP*, pp. 249-252, 1992.

[2] Benallal, A., and Gilloire, A., "Instabilité et stabilité numérique des algorithmes des moindres carrés transversaux rapides excités par la parole", *GRETSI*, pp. 509–512, 1989 (in French).

Noise Type		Performance (%)
White	Noisy	44.20
	Denoisy	71.40
Pink	Noisy	33.94
	Denoisy	61.20
Babble	Noisy	50.99
	Denoisy	81.59
Factory	Noisy	37.39
	Denoisy	73.02

Table 5: GMM Based SI performance with noisy and denoisy data generated from Spidre Corpus

[3] Dempster, A., Laird, N. and Rubin, D., "Maximum Likelihood from Incomplete Data Via the EM Algorithm", *J. Roy. Stat. Soc,*, Vol. 39, pp. 1-38, 1977.

[4] Furui, S., "An Overview of Speaker Technology", *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, pp. 1-9, 1994.

[5] Gabrea, M., Mandridake, E., Menez, M., and Najim, M., "Two Microphones Speech Enhancement System Based on a Double Fast Recursive Least Squares (DFRLS) Algorithm", *EUSIPCO*, pp. 983–986, 1996.

[6] Reynolds, D. A., "The Effect of Handset Variability on Speaker Recognition Performance: Experiments on the Switchboard Corpus", *ICASSP*, pp. 113-116, 1996.

[7] Tadj, C. , Dumouchel, P. , Mihoubi, M. et Ouellet, P., "Environment Adaptation and Long Term Parameters in Speaker Identification", *European Speech Communication Association (EuroSpeech)*, Budapest, Hungary, septembre 5-9, 1999.

[8] Tadj, C., "Features Extraction for Speaker Identification", Information Systems, *Analysis and Synthesis / Systemics, Cybernetics and Informatics (ISAS/SCI2000)*, Orlando, Florida, 23-26 juillet, 2000.

[9] Widrow, B., and al, "Adaptive Noise Cancelling: Principles and Applications", *Proc. IEEE*, vol. 63, no. 3, pp. 1692–1716, 1975.