6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

A ROBUST ISOLATED WORD RECOGNIZER FOR HIGHLY NON-STATIONARY ENVIRONMENTS. RECOGNITION RESULTS

A. Álvarez, R. Martínez, P. Gómez, V. Nieto and, Mª. M. Pérez

Departamento de Arquitectura y Tecnología de Sistemas Informáticos, Facultad de Informática
Universidad Politécnica de Madrid, Campus de Montegancedo, s/n
Boadilla del Monte, 28660, Madrid, SPAIN
e-mail: pedro@pino.datsi.fi.upm.es

ABSTRACT

Through the present paper, the evaluation results of a *Speaker Independent Robust-to-Noise Isolated Word Recognizer* are presented. The system, which is in part the results achieved by the project *IVORY* (ESPRIT project No. 20277) [6], is intended for working in highly non-stationary environments. The system comprises two main modules: the *Noise Cancellator* and the *Speech Recognizer* itself. System robustness is achieved in the noise cancellation module. This module incorporates an *Adaptive Filter* [7] [13], operating in the time domain, and a subsequent *Spectral Subtraction* step [2] operating in the frequency domain with the enhanced signal provided by the previous stage. Recognition results for different noise cancellation configurations and for several *Parameter Extraction Front-Ends*, including LPC [7], FFT Cepstrum [3] and PLP based methods [8] are presented.

Keywords: robust speech recognition, non-stationary environments, recognition results.

1. INTRODUCTION

Environmental noise is one of the major problems to be faced by speech recognition systems working under real conditions, as most systems performance degrade substantially with low S/N ratios [1]. This is particularly important in non-stationary environments as in cars, plane cockpits, workshops with heavy machinery, crowded public places, etc. These scenarios are characterized by high noise levels (of about 95 dB with SNR´s about 0- 10 dB) produced by different sources and where noise contributions are especially harsh, as many different noise sources may be active simultaneously, varying dramatically with time. To face with the treatment of noise, *Robust Speech Recognition* has become an objective in itself [4].

Along the years, several independent techniques have been proposed. Most of them are based in two basic approaches: *Adaptive Time-Domain Filtering* (ATDF) [14], and *Frequency-Domain Spectral Subtraction* (FDSS) [2]. Both methods have been used independently with different degree of success, depending on the problem being solved. Generally, ATDF has been used with array microphones, ranging from two-microphone to multiple-microphone systems. On the other hand, FDSS has been preferred for systems where a single microphone is a requirement (e.g. Telephone speech recognition). This technique has found its niche in applications where noise is more or less a quasi-stationary and evenly distributed process. ATDF has the advantage of being highly suitable to deal with non-stationary signals as speech and most kinds of noise, although at a higher cost (both in instrumentation and in computational power). Although both methods can be integrated in the same system, most of the times they have been used separately [5]. However, the co-operation of both procedures allows important gains in the S/N ratio to be achieved, improving the overall recognition results, as well.

2. HYBRID NOISE CANCELLATOR

The hybrid noise-cancellation scheme proposed, which can be seen in Fig. 1, is based in a two-microphone array (Speech Source or *primary*, and Noise Source or *reference*) [10]. The primary microphone is placed close to the speaker, and the second one, at a certain distance in order to acquire a good estimation of the noise received by the primary microphone at the same time it avoids capturing speech.

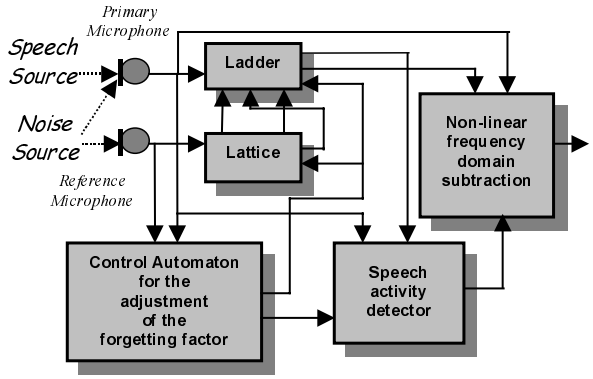

Figure 1. Hybrid Speech Enhancement system used throughout the experiments described [5].

Regarding the working requirements, the cancellation scheme assumes:

- It should deal with noise levels over 85 dB and SNR´s about 0- 10 dB.
- Local SNR at some frequencies can be negative.
- Noise has continuously changing characteristics.

- The noise spectrum is quite spread over the speech zone and it can contain in itself speech sounds (*cocktail party effect*).

2.1 Time Domain Filter

In our system, the adaptive processing is achieved by a *Time-Domain Joint-Process Estimator* implemented as a *Time-Domain Lattice-Ladder Filter* with a least-squares estimation algorithm (*Recursive Least Squares Lattice* using *a posteriori* estimation errors) [9]. This filter has two inputs: the lattice part is fed with the noise source (reference), and the primary signal is inserted through the ladder part. The lattice is therefore adapted with the noise characteristics and a set of backward prediction errors is attained. The main advantage of this method is that no matter the noise level is, if the reference signal is good enough, a considerable amount of cancellation is achieved. Nevertheless, one of its major limitations is the computational complexity. The longer the filter, the higher the cancellation gain, but the number of operations grows accordingly.

The *forgetting factor* adjustment, controlled by a finite automaton monitoring the power of incoming primary and reference signals, guarantees the stability of the filter against sudden changes in power ratio between the two channels [11]. This mechanism also provides an effective way to minimise the required locking period of the filter after the production of an utterance.

2.2 Spectral Domain Filter

The frequency-domain filter here proposed is very efficient, and produces a level of cancellation of the same order, or higher than the adaptive filtering. In this case it is used as a previous processing block for obtaining an enhanced version of the signal and a first estimation for speech detection [12]. Its major limitation is that it requires a sufficient SNR (which is not always possible). Otherwise, two effects will appear: firstly, spectral lines completely buried into noise (local negative SNR) will be removed, and secondly it will be very difficult to make a decision on the presence of speech (especially if we consider that the noise in itself can contain speech sounds). The determination of non-speech periods is imperative to make an accurate comparison of both channels, as this has to be carried out when speech is not present.

The objective of the processing is to calculate the ratio between every spectral line of the *Joint Process Error* (JPEr) and the *Joint Process Estimate* (JPEs), and modify the JPEr accordingly to produce an estimate of the noise. To implement this filtering, the JPEr output of the lattice-ladder filter is used as the primary signal, and a noise estimate obtained from the JPEs output is used as the reference. These two signals are segmented in overlapped windows and transformed into the frequency domain using the short-time Discrete Fourier transform.

The maximum of two consecutive values, calculated from a smoothed and weighted (using a logarithmic law) version of that reference channel, is selected. Later, that maximum is subtracted from the primary input in the frequency domain to produce the enhanced speech trace for that frequency value. Therefore, if speech is not present or there are not speech contents at that frequency, a larger cancellation gain is obtained. On the other hand, spectral components of speech, whose levels are now higher than noise, will remain almost unaffected. After the subtraction step, a half wave rectification is carried out to avoid the possibility of obtaining negative values for the norm of the vector [2]. Finally, the phase of the enhanced signal is recovered from the JPEr trace.

3. RECOGNITION FRAMEWORK

The structure of the *Speaker Independent* Isolated Word *Speech Recognizer* used for the experiments was the following:

- Noise cancellator as a preprocessing step.
- Feature extraction with three different schemes.
- Vector quantizer using 256 indexes.
- Word model parser using DHMM.

The schemes devoted to the parameter extraction step was:

- LPC based. The *Gradient Adaptive Lattice* algorithm, Levinson-Durbin Routine and the Transfer Function [7] are used to extract 20 energy bands from the spectrogram. Delta bands with a 5-frame interval are also calculated.
- FFT based [3]. In this case 10 MFCC+ 10 Δ-MFCC are calculated. The number of frames involved in the calculation of the Δ-parameter was 5.
- PLP based [8]. Using a normal configuration 30 parameters (10+10 Δ+ 10 Δ^2) are extracted.

Regarding the recording conditions, the framework was the following:

- Noise was induced in the environment during the recordings using a noise source consisting in a power loudspeaker fed with a car-racing game noise. Noise levels were measured at the primary microphone with a standard sonometer. The primary microphone was at a distance of 1.5 m from the noise source. The level of noise ranged from 65 dB to 95 dB. The speaker's loudness was kept around 80-85 dB. Stress was induced in silent recordings used for the training set.
- The methodology for adding noise was identical to the conditions found in an Arcade Game environment. Noise was present in the microphones at the time speech was recorded to generate the noisy set of words for testing. These conditions are the closest to the real situation.
- The *Speech Database* was produced including the 30 words listed in Table 2. The number of recordings per word was 1 for clean, 1 for stressed speech and 1

for noisy speech. Clean and stressed speech were used for model training and noisy speech is only used for testing purposes.

- The speakers were equally distributed between both sexes. The language for the database was English but the dialectal variants included comprised only Spanish-native speakers. The speakers' age ranged between 18 and 50 years, the average age being 22.5.
- For the experiments to be referred 16 speakers were used in training, equally distributed between sexes. In testing, a group of different 24 speakers was used (12 Spanish-native male and 12 Spanish-native female).
- The set of words used in the experiments was the total set of 30 words as listed in Table 1.
- The instrumentation used in the recording was a pair of cardiodid microphones and a high quality 16-bit resolution sound card. The recordings were made at a sample frequency of 22,050 Hz and afterwards re-sampled to 11,025.

Double	Four	One	Left	Split	Turn
Down	Go	On	Next	Stand	Two
Eight	Hit	Right	Nine	Stop	Up
End	Jump	Seven	No	Ten	Yes
Five	Last	Six	Off	Three	Zero

Table 1. List of words used in the Speech Recognition experiments.

4. RESULTS AND DISCUSSION

A set of experiments has been conducted using the speech database described previously. The training material consisted in clean and stressed speech. In some of the experiments, filtered speech coming from the *Noise Cancellator* was also used. The test sets included clean utterances, stressed speech, noisy speech, filtered speech with the ATDF process alone (A) and filtered speech produced by the completed speech enhancement system (B).

The differences among the power of the speech signal before and after being filtered power is shown in Figure 2. An improvement of more than 20 dB in the SNR can be observed between the noisy speech trace and the resulting enhanced speech.

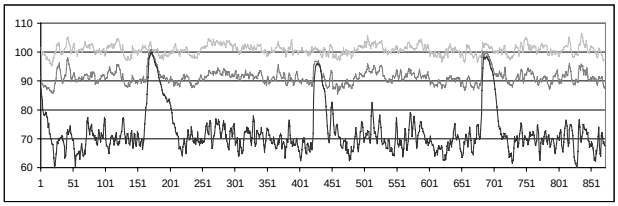

Figure 2. Energy corresponding to the noisy speech trace, the cleaned speech trace after de ATDF process and, the output of the frequency domain filter (FDSS).

The accuracy of recognition for the experiments is given using the common expression for the Word Error Rate:

$$WER = 100 \cdot \frac{N_{insertions} + N_{deletetions} + N_{substitutions}}{N} \quad (1)$$

In isolated word recognition, the insertion errors are produced by false starts and hesitations. Deletion errors have their origin in end-point detection errors.

Figure 3, Figure 4 and Figure 5 show the results for the different feature extraction methods. The results for the quiet and Lombard utterances are far from the expected ones. We find two main reasons:

- The simplicity of the speech recognition framework and the small size of the database are not enough for capturing all the necessary information for achieving the desired results.
- The speaker variability introduced by English non-native speakers accounts rather negatively in a speaker independent system. In fact, the recognition results attained for the same group of speakers using a 30 word set from Spanish is around 94.5% when the system configuration and the training and test set sizes are kept equally.

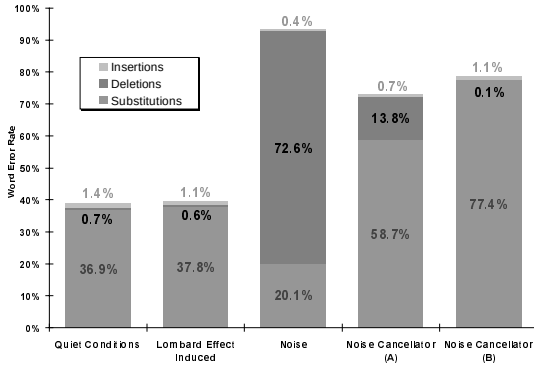

Figure *3*. Recognition results for the MFCC front-end.

In the noisy tests, the main source of degradation comes from errors in the word-boundary detection process. This situation is clearly alleviated when the Noise Cancellation stage is applied. However, there is a drawback effect on that action. The introduction of the cancellation scheme into the recognizer produces a significant increase in the substitution error rate.

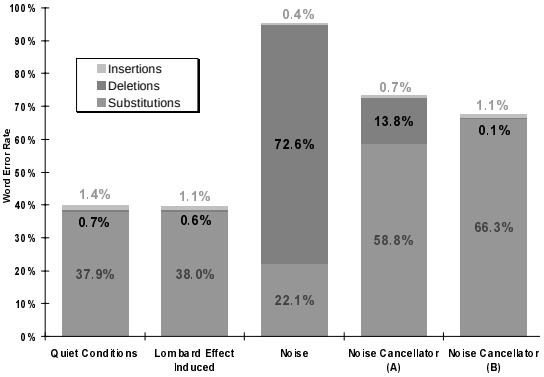

Figure 4. Recognition results for the LPC based front-end.

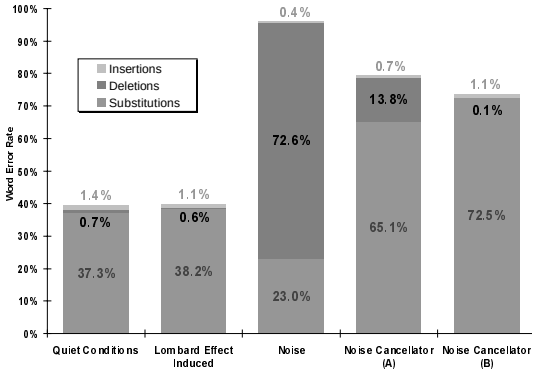

Figure 5. Recognition results for the PLP based front-end.

The comparison among the results produced by the three different spectral-analysis methods reveals a similar behavior. However, the LPC based method behaves a little better when the cancellation scheme is applied. Finally, the incorporation of filtered utterances in the training group improves the recognition rates as can be seen in Figure 6. However, that improvement is still far from the silent-condition situation.

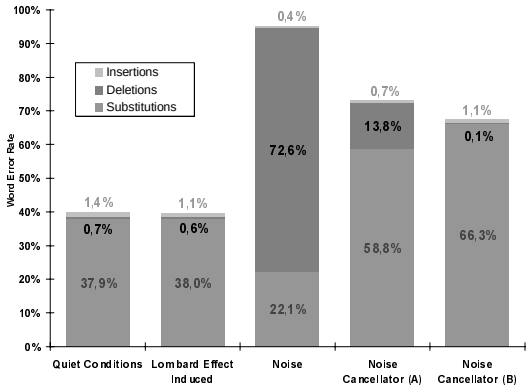

Figure 6. Recognition results for the MFCC front-end training also with the filtered utterances from the training group.

5. CONCLUSIONS

Time-Domain Adaptive Filtering and *Frequency Domain Adaptive Spectral Subtraction* may be combined for speech enhancement. The gain in the S/N ratio improves substantially the operation of the associated *Speech Recognition System*, especially under heavy and non-stationary noise conditions. The insertion of this kind of preprocessing stage implies dealing with a new source of intra-speaker variability, as the noise removal procedure introduce slight changes in the frequency contents of the spectrogram.

6. ACKNOWLEDGEMENTS

This work is being funded by grants TIC96-1889-C, TIC97-1011, from the Comisión Interministerial de Ciencia y Tecnología and by an Agreement between UPM and the Centre Suisse d'Electronique et de Microtechnique.

7. REFERENCES

[1] Agaiby, H., et al., "Commercial Speech Recognizers Performance under Adverse Conditions, A Survey", *Proc. of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-à-Mousson, Francia, 17-18 April 1997, pp. 163-166.

[2] Boll, S. F., "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. on ASSP*, vol. ASSP-27, No. 2, April 1979, pp. 113-117.

[3] Davis, S. B. and Mermelstein, P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. ASSP-28, no. 4, August 1980, pp. 357-366.

[4] Furui, S., "Recent Advances in Robust Speech Recognition", *Proc. of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Rec. for Unknown Comm. Channels*, Pont-à-Mousson, France, 17-18 April 1997, pp. 11-20.

[5] Gómez, P., et al., "A Hybrid Signal Enhancement Method for Robust Speech Recognition", *Proc. of the Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, Tampere, Finland, May 25-26, 1999 (to be published).

[6] IVORY project, http://tamarisco.datsi.fi.upm.es/ PROJECTS /IVORY/ivory.html.

[7] Haykin, S., *Adaptive Filter Theory*, 3rd Ed., Prentice Hall, Englewood Cliffs, N. J., 1996.

[8] Hermansky, H., "Perceptual linear predictive (PLP) analysis of speech", *Journal of Acoustic Society of America*, Vol. 87, no. 4, April 1990, pp. 1738-1752.

[9] Martínez, R., et al., "ASR in Highly Non-Stationary Environments using Adaptive Noise Cancelling Techniques", *Proc. of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Rec. for Unknown Comm. Channels*, Pont-à-Mousson, France, 17-18 April, 1997, pp. 181-184.

[10] Martínez, R. Et al., "Implementation of an Adaptive Noise Canceller on the TMS320C31-50 for Non-Stationary Environments", *Proc. of the 13th International Conference on Digital Signal Processing*, Santorini, Greece, 2-4 July 1997, pp. 49-52.

[11] Martínez, R., et al., "Dynamic Adjustment of the Forgetting Factor in Adaptive Filters for Non-Stationary Noise Cancellation in Speech", *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP´98*, Seattle, Washington, USA, May 12-15, 1998, Vol. 2, pp. 1009-1012.

[12] Martínez, R., et al., "Combining Linear and Non-Linear Processing in the Time and in the Spectral Domain for Non-Stationary Noise Filtering", *Proc. of the 1999 IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP´99*, Antalya, Turkey, 20-23 June 1999 (to be published).

[13] Proakis, J. G., *Digital Communications*, 2nd. Ed, McGraw Hill, 1989.

[14] Widrow, B., et al., "Adaptive Noise Cancelling: Principles and Applications", *Proc. of the IEEE*, Vol. 63, No. 12, December 1975, pp. 1692-1716.