COMPARISON OF SEVERAL PREPROCESSING TECHNIQUES FOR ROBUST SPEECH RECOGNITION OVER BOTH PSN AND GSM NETWORKS

Chafic Mokbel, Laurent Mauuary, Denis Jouvet and Jean Monné
France Télécom - CNET / LAA / TSS / RCP
2 av. Pierre Marzin, 22307 Lannion cedex, France
e-mail: mokbel(jouvet, monne)@lannion.cnet.fr

ABSTRACT

In this paper several preprocessing techniques used to improve speech recognition performance are compared over both PSN and GSM networks. Recognition experiments are conducted on a digit database in a speaker-independent isolated-word mode in order to evaluate the performances under within- and cross-network (PSN and GSM) conditions. Two classes of preprocessing techniques are distinguished depending on whether they deal with additive ambient noise or convolved perturbations. The first class preprocessing techniques are based on spectral subtraction. In the second class, the low frequencies of cepstral trajectories are eliminated in order to reduce convolved disturbances. Blind equalization adaptive filtering has been proposed to reduce channel effects. In this study, channel equalization and speech enhancement techniques are combined and compared. Different recording conditions may be integrated in order to increase robustness. This is done during the training phase using HMM models with variable parameters. Recognition results are analysed as a function of recording conditions.

1. INTRODUCTION

Successful speech recognition systems should be robust to changes in the ambient and transmission conditions of the observed speech signal. With the rapid development of mobile communication systems, cellular networks open a wide range of applications for speech recognition. This new environment and, more specifically, the GSM network differ from the classical PSN environment in both the characteristics of the transmission channel and of the user environment (calling from running cars, outdoors, ...). It is therefore necessary to find a preprocessing technique that makes a recognition system more robust to changes in the ambient and transmission environments.

Spectral subtraction-type speech enhancement techniques are generally used to reduce the additive ambient noise [6]. Telephone channel effects are often associated with low frequencies in cepstral (or log-spectral) trajectories [4][2]. Existing techniques usually reduce the low frequencies of cepstral trajectories in order to filter out the convolved disturbances in the observed signal. The use of cepstral subtraction [4], highpass filtering of cepstral trajectories (RASTA) [2][3][5] and blind equalization using adaptive filtering [7] on PSN telephone speech data has resulted in large improvements. In this paper, speech enhancement and channel equalization techniques are compared and combined. The aim is to increase the robustness of the speech recognition system in within- and cross- PSN and GSM networks.

The robustness of a speech recognition system based on Markov modelling is increased by the inclusion, in the training set, of data recorded under various conditions over both PSN and GSM environments. It is well known that during the training phase the stochastic Hidden Markov Models (HMM) mix up all the conditions present in the data. Moreover, the modelling of the conditions is improved by increasing the range of HMM parameters, by using mixtures of Gaussian densities.

This paper addresses the issues of preprocessing techniques, as well as the choice of the training data and the effect of increasing the range of HMM parameters. In section 2 speech enhancement techniques based on spectral subtraction are presented. Channel equalization algorithms, used to increase robustness of speech recognition systems over telephone networks, are described in section 3. A method allowing to combine speech enhancement and channel equalization techniques is proposed in section 4. Section 5 addresses the HMM modelling issues. The characteristics of the digit database used in our recognition experiments are detailed in section 6. Recognition experiments and the obtained results in several conditions are reported and discussed in section 7. Finally, section 8 draws the main conclusions.

2. SPEECH ENHANCEMENT TECHNIQUES

Speech enhancement techniques improve recognition performances when stationary additive ambient noise is present, as in the car environment [6]. In this paper, we consider the recognition over the telephone network with the constraint of a single sensor to capture the speech waveform - the microphone of the telephone handset. One-sensor speech enhancement techniques should therefore be used, namely spectral subtraction [1]. With spectral subtraction method, the spectral density of stationary additive noise is estimated during the nonspeech periods. This estimate is then subtracted from the short term spectral densities of the observed signal. The phase of the noisy frames helps to compute the

enhanced signal frames. The musical noise observed in the enhanced signal is the main problem of this method. Several variants were proposed in the literature to reduce this musical noise [1][6]. In this study, the basic spectral subtraction method is investigated to improve the robustness of speech recognition in noisy environment.

To fully understand the spectral subtraction method, consider that $x(t)$ is the observed noisy signal corresponding to the clean signal $s(t)$ and the stationary additive ambient noise $n(t)$:

$$x(t) = s(t) + n(t) \quad (1)$$

If we suppose that each signal has a zero mean and that $n(t)$ and $s(t)$ are not correlated, (Eq. 1) becomes in the autocorrelation domain:

$$R_{xx}(\tau) = R_{ss}(\tau) + R_{nn}(\tau) \quad (2)$$

The autocorrelation function is directly related to the short-term spectral density, and Eq. 2 becomes:

$$\Gamma_{xx}(f) = \Gamma_{ss}(f) + \Gamma_{nn}(f) \quad (3)$$

For a given frame, $\Gamma_{xx}(f)$ is observed. Thus, an optimal estimate of $\Gamma_{ss}(f)$ may be found by:

$$\overline{\Gamma_{ss}(f)} = \Gamma_{xx}(f) - \overline{\Gamma_{nn}(f)} \quad (4)$$

where $\overline{\Gamma_{nn}(f)}$ is estimated during the nonspeech periods.

3. CHANNEL EQUALIZATION TECHNIQUES

The effects of the PSN telephone channel have been studied in earlier works [4][5]. The telephone handset microphone actually captures $x(t)$, a mixture of clean speech waveform and of ambient noise waveform. If we assume that the channel (microphone, telephone network, ...) acts as a linear time invariant (LTI) filter $h(t)$, the observed signal $y(t)$ can be written:

$$y(t) = x(t) \otimes h(t) = \left[s(t) + n(t) \right] \otimes h(t) \quad (5)$$

Reducing the channel effects is a blind equalization problem as only a single-sensor signal is observed at the input of the recognition system. From Eq. 5, telephone channel effects appear in the feature vector space (MFCC cepstral space in our system) as additive bias, slowly varying with time. With the use of data collected from many different conditions, the HMM models broaden their Gaussian densities in order to consider the bias in the observed feature vectors. This reduces their discriminating capacities [5]. The robustness of the recognition system is increased by the removal of the low frequencies in the cepstral trajectories. Three algorithms which perform channel equalization are described as follows:

3.1. Cepstral Normalization

Cepstral normalization [4] consists in estimating the long-term cepstrum of speech for a given call and subtracting it from the cepstra of the frames. It has been shown [7] that a long-term cepstrum estimated on few seconds of speech produces a reliable estimate of the channel effect. The long-term cepstrum estimated on the nonspeech parts of the signal includes the stationary silence cepstrum.

3.2. Highpass filtering of cepstral trajectories

One possible way to reduce the low frequencies of cepstral trajectories is the application of a highpass filter on the cepstral trajectories [2][3][5]. The response time of the highpass filter should be large enough to reduce the low frequencies in a long-term sense. If the response time of the highpass filter is limited to the duration of a stationary segment of speech, the filter will inverse the local segments characteristics. IIR first order filters were used in our experiments ($\lambda = 0.04$):

$$\frac{1 - z^{-1}}{1 - (1 - \lambda) \cdot z^{-1}} \quad (6)$$

3.3. Blind equalization using adaptive filtering

The reduction of the channel effects in the single-sensor observed signal is a blind equalization problem. In digital communication blind equalization schemes adapt the parameters of the equalizing filters on the basis of the decision module and some known characteristics of the conveyed signal. The same principle may be applied for implementing an adaptive filter [8] in order to equalize the channel effects in the observed speech signal.

As no reference signal is available, the adaptive scheme should make use of some a priori statistics on the speech signal in order to perform equalization. A possible criterion is to consider that in the long-term sense the spectral density of the speech signal is constant. An error is then computed as the difference between the spectral density of the signal at the output of the adaptive filter and the constant reference spectral density. The filter parameters can be adapted on the basis of this error.

The energies at the output of a MEL frequency filterbank are computed in the feature extraction module of the recognition system. This induces a frequency implementation (esp. "Circular-Convolution method" [8]) of the LMS adaptive filter. Fig. 1 shows the implemented adaptive filter.

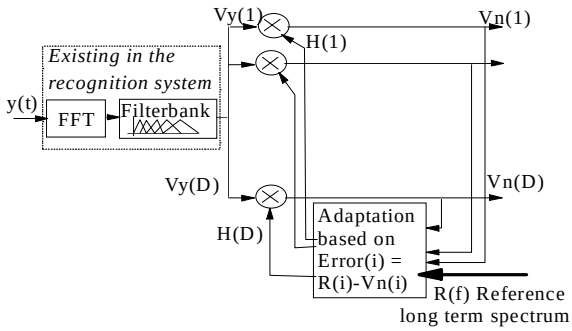

Fig. 1: *LMS adaptive filter for blind equalization.*

It has been shown [7] that for speech signals the proposed filter converges to the optimal equalizing filter

if the channel effect is LTI. This adaptive scheme is based on long-term statistics. The power of the reference signal is modulated by the power of the input frame to guarantee a correct convergence of the filter. The power of the input speech signal varies with the pronounced sound and, if the power of the reference spectrum is kept constant, this makes the filter inverse the spectrum of the most energetic sounds at convergence.

4. COMBINING CHANNEL EQUALIZATION AND SPEECH ENHANCEMENT

In Eq. 5, the observed signal is the sum of the clean signal and the additive ambient noise, convolved with a linear filter slowly varying with time.

$$y(t) = [s(t) + n(t)] \otimes h(t) = s(t) \otimes h(t) + n(t) \otimes h(t) \quad (7)$$

In order to apply spectral subtraction the additive noise to be removed must be stationary and noncorrelated with the desired signal. As the additive noise n(t) in Eq. 7 is considered to be stationary and noncorrelated with the clean signal s(t), and supposing h(t) to be a LTI filter, the signal $n(t) \otimes h(t)$ in the sum of Eq. 7 is stationary and noncorrelated with $s(t) \otimes h(t)$. Based on this argument, spectral subtraction should be applied first when combining speech enhancement and channel equalization. It is fortunately not difficult to apply spectral subtraction first, especially when we consider cepstral subtraction or highpass filtering of cepstral trajectories as the channel equalization technique.

5. HMM MODELLING ISSUES

A more robust recognition model may be built to improve recognition performances by the integration of a number of recognition conditions. This could be done by collecting training speech data in a number of conditions. However, the variability of the training speech data decreases the discriminating nature of the HMM models for a fixed number of parameters. In order to overcome this limitation the number of parameters may be increased by the use mixtures of Gaussian densities. The CNET HMM system, PHIL90, is used in our experiments. 30 states word models represent the digits. Silence models were placed at both sides of the digit models to avoid a precise endpoints detection.

6. DATABASE

A French digit database is used in the recognition experiments. This database is collected over two telephone networks: PSN and GSM. Speakers calling from different regions of France repeat the 10 digits. Each call corresponds to a given speaker. Nearly 1000 calls are collected in the PSN environment and about 1300 calls in the GSM environment. The digits are automatically detected using a speech/nonspeech detector. With the help of a human listener, only the well detected digits are kept in the database.

Several recording conditions are present in this database. Local and long-distance calls conditions are present in the PSN part of the database. Outdoors, indoors, calls in stopped and running cars are included in the GSM part of the database.

7. RECOGNITION EXPERIMENTS

Recognition experiments are carried out with training in one network environment and testing in both network environment. The combination of the two network environments in the training set is also tested. Each of PSN and GSM databases were divided into two parts; one half for training and the other for testing. Two recognition experiments are performed for each training condition: PSN, GSM or joint PSN and GSM. Each time one half of the database is used in the training phase. The average of the obtained error rates are given.

The first set of experiments investigates the improvements obtained with the channel equalization techniques: cepstral normalization, highpass filtering and blind equalization. Fig. 2 gives the obtained error rates when mixture of 8 Gaussian densities are used in HMM modelling. Cepstral normalization produces the better performances in all the conditions. However, this technique is off-line. Blind equalization using adaptive filtering, an on-line method, produces larger improvements than highpass filtering.

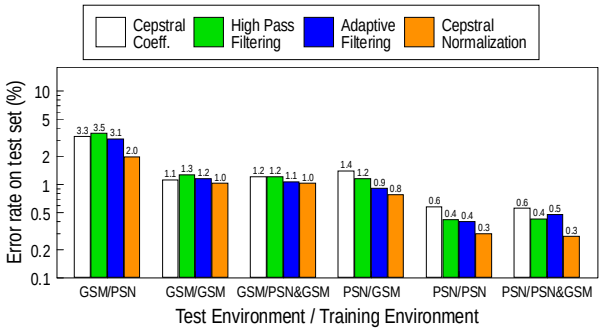

Fig. 2: *Channel equalization methods.*

Fig. 3 shows the recognition performances after the introduction of spectral subtraction. The spectral subtraction technique, when used alone, only improves the recognition performances when training is performed in the PSN environment and testing in the GSM environment. The combination of spectral subtraction and blind equalization using adaptive filtering produces large improvements, especially when training is performed in the PSN environment. The performances deteriorate slightly when spectral subtraction is used in the GSM training condition. This is probably due to existing impulsive GSM noises that are reinforced, with spectral subtraction, in the training signals.

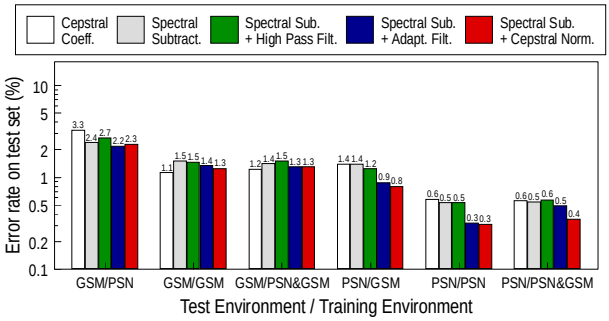

Fig. 3: *The introduction of spectral subtraction in the preprocessing scheme.*

Fig. 4 shows the performances obtained with various preprocessing techniques function of the ambient condition (training in the PSN environment). The long-distance calls in the PSN environment present, as expected, more problems than the local calls. The outdoors condition is surprisingly the most difficult, even in comparison with the running vehicle condition. It appears also that the stopped vehicle condition is equivalent to the indoors condition. With reference to figure 4, the combination of spectral subtraction and blind equalization helps to increase the robustness in almost all conditions. Spectral subtraction is important for running car and outdoors conditions, due to low SNR in these conditions.

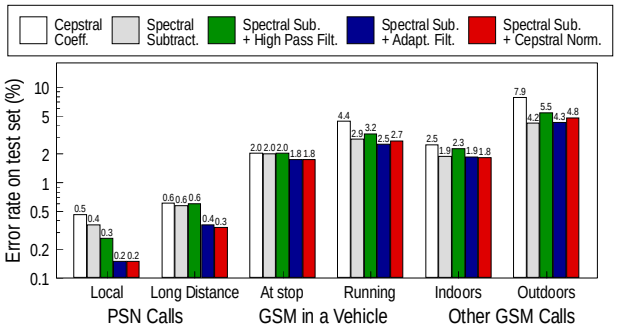

Fig. 4: *Recognition error rates in various conditions.*

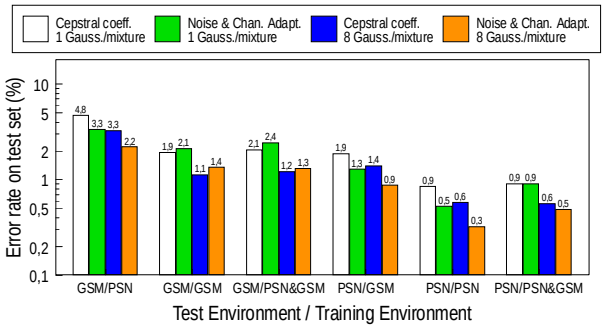

Fig. 5: *Combining modelling issues with preprocessing techniques.*

Fig. 5 shows the comparison between HMM with single Gaussian densities to HMM with mixtures of 8 Gaussian densities. The effect of preprocessing is also investigated. We can conclude from these results that modelling with more parameters and the integration of several training conditions does improve the robustness. The improvements obtained using preprocessing techniques can be added to those which are relative to modelling with more parameters. Unfortunately, the performances deteriorate when GSM data are present in the training set and when spectral subtraction is used. We believe that this is due to the fact that GSM impulsive noises are amplified with spectral subtraction and they are poorly handled by the HMM modelling.

8. CONCLUSIONS

This paper addresses the problem of robust automatic speech recognition in within- and cross- PSN and GSM networks. It has been shown that the worst conditions are long-distance calls for PSN environment and outdoor and running car conditions for GSM environment.

Speech enhancement and channel equalization increases the robustness of the system. The combination of spectral subtraction and blind equalization based on adaptive filtering leads to a efficient preprocessing module, especially when training is performed only in the PSN environment.

We also investigated the inclusion of training data recorded in various conditions and the use of HMM with more parameters in order to increase robustness. It has been shown that HMM, with mixture of Gaussian densities, trained on both PSN and GSM data provides robust recognition systems. The experiments show that the robustness can be increased by the combination of preprocessing methods and robust HMM modelling.

References

[1] M. Berouti, R. Schwartz and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," Proc. ICASSP'79, p. 208-211, 1979.

[2] H. Hermansky, N. Morgan, A. Bayya & P. Kohn, "Compensation for the Effect of the Communication Channel in Auditory-like Analysis of Speech (RASTA-PLP)," Proc. EuroSpeech'91, p. 1367-1370, 1991.

[3] H.G. Hirsch, P. Meyer & H. Ruehl, "Improved Speech Recognition Using High-Pass Filtering of Subband Envelopes," Proc. EuroSpeech'91, p. 413-416. 1991.

[4] C. Mokbel, J. Monné & D. Jouvet, "On-Line Adaptation of a Speech Recognizer to Variations in Telephone Line Conditions," *Proc. EuroSpeech'93*, p.1247-1250, 1993.

[5] C. Mokbel, P. Pachès-Leal, D. Jouvet & J. Monné, "Compensation of Telephone Line Effects for Robust Speech Recognition," Proc. ICSLP'94, p. 987-990, 1994.

[6] C. Mokbel & G. Chollet, "Automatic Word Recognition in Cars," *IEEE Trans. SAP*, Vol. 3, n° 5, Spet. 1995, p. 346- 356.

[7] C. Mokbel, D. Jouvet & J. Monné, "Blind Equalization using Adaptive Filtering for Improving Speech Recognition over Telephone," Proc. EuroSpeech'95, p. 1987-1990, 1995.

[8] J.J. Shynk, "Frequency-Domain and Multirate Adaptive Filtering," IEEE Signal Proc. Magazine, p. 15-37, Jan. 1992.