CELLULAR PHONE SPEECH RECOGNITION: NOISE COMPENSATION vs. ROBUST ARCHITECTURES

J-B. Puel and R. André-Obrecht
IRIT - Université Paul Sabatier
118, route de Narbonne
31062 Toulouse cedex - France
{puel, obrecht}@irit.fr

ABSTRACT

This paper addresses the problem of speech recognition through telephonic networks. When the communication channel is unknown, the important mismatch between training data and signal encountered in recognition phase decreases drastically the performances of the recognition systems. In this context, we compare a classical approach: the noise compensation method with novel robust networks modellings aiming to incorporate and manage more variability in the training data.
We introduce multi-HMMs and multi-transitions systems, trained with data recorded through analog switched network and cellular phone network. These architectures present best results and succeed in improving the recognizers robustness since they achieve up to 77 % reduction of the error rate for a system trained for switched telephonic network and used with cellular phone. Nevertheless, this modelling requires training data recorded in both environments; when such data are not available, noise cancellation or channel compensation are the only affordable solutions.

1. INTRODUCTION

Hidden Markov Model-based speech recognizers offer realistic solutions for creating interactive vocal servers, but an acoustical mismatch between training and testing conditions causes a severe degradation in the recognition performances. When using the telephonic channel, it is impossible to know in advance in which noise environment the system will be used, and what kind of telephonic network will support the communication. To minimize this effect, many studies have been performed to find more robust signal processings (as spectral subtraction [2]) and efficient distance measures [6] or to propose or adapt new recognizer models [4].
In this paper, we propose and compare two HMM-based speech recognizers :

- a **Noise Compensation system** where an adaptative noise spectral subtraction is performed; its originality lies upon the use of a segmentation algorithm with a speech activity detector,
- a **robust HMM architecture** : multi-network and multi-transitions modelling aiming to support and manage more variability during the training of the system.

We compare the results obtained by these methods on CNET speech corpora recorded with telephonic and cellular phone noise conditions (GSM system).

2. NOISE COMPENSATION SYSTEM

2.1. General method

The first set of methods we introduce consists in preprocessing learning and test data in order to deal with the noise that corrupts the speech signal.
When assuming that the signal $x(n)$ is the sum of the speech $s(n)$ and a non correlated noise $b(n)$, short term stationary

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

In the spectral domain

$$\Gamma_x(\omega) = \Gamma_s(\omega) + \Gamma_b(\omega) \quad (2)$$

where $\Gamma_x(\omega)$ is the short term power spectral density of $x(n)$. The general method of spectral subtraction [2] consists in computing the speech spectrum estimate

$$\hat{S}(\omega) = |X(\omega)| - |\hat{B}(\omega)| \quad (3)$$

The principal problem is the estimation of the noise $|\hat{B}(\omega)|$, in this purpose we proceed the following steps :

- an automatic segmentation algorithm [1] detects the frontiers of quasi-stationary zones of the signal,

- monitoring of the curvilinear abscissa of the temporal signal [8] gives a crude detection of speech endpoints. The temporal coordination between these results provides the "noise/speech" labelling of each segment and robust endpoints,

- evaluation of $|\hat{B}(\omega)|$ on the longest noise segment.

This mean vector is subtracted to each frame in the spectral domain, using a weight coefficient function of the SNR.

2.2. Segmentation

The segmentation algorithm is the "Forward-Backward Divergence" [1] that locates quasi-stationary zones in the signal. This first step of the processing will bring information for the speech activity detection, and make easier the pseudo-diphone modellisation used by our recognition system.
We assume the signal is represented by a string of homogeneous units, each of them represented by an Auto-regressive model. The method consists in detecting changes in the model parameters. We obtain segments of three kinds :

- stationary segments corresponding to steady parts of the signal,

- transition segments in which a formantic structure can be found, with monotonous behavior,

- short segments (10 ms) corresponding to articulatory changes, like plosives explosion.

This segmentation algorithm detects all speech/non-speech boundaries, but does not identify them as such.

2.3. Noise estimation

2.3.1. Noise/Speech detection

The segmentation module provides a list of boundaries corresponding to spectral modifications of the signal, but without identifying the location of each segment : speech zone or a noise zone. Many speech activity detectors have been presented yet, but none use a segmentation algorithm; some techniques use pattern recognition methods, like [5], some use particular acoustic coefficients like zero-crossing rate, energy, etc... Our approach is based on the fact that, even if energy is not a robust parameter in noisy conditions, maxima of the signal amplitude always correspond to vocalic cores [8].
So, we process the two following steps :

Static labelling : In a first time, the curvilinear abscissa $s(t)$ of the speech signal $y(t)$, where t is the sample index, is computed. Let the function :

$$S(n) = s(nL) - s((n-1)L)$$

where L is a fixed number of samples (a frame).
$S(n)$ represents a mean value of the "curve length" by time units. Assuming the noise to be stationary for each segment, the S function will show very little variation in noise zones, increase perceptibly when entering a speech zone and decrease when coming back to a noise zone.
The mean $\overline{S_i}$ and standard deviation $\sigma(S_i)$ of S for the segment i will represent our speech or noise indicator.

Two thresholds are used : λ_1 and λ_2, automatically computed on the first signal frames assumed to be noise only.

$$\lambda_1 = \overline{S_b} + \sigma(S_b) \text{ and } \lambda_2 = n \times \lambda_1$$

where n is a threshold to discriminate the mean level of curvilinear abscissa between noise and noisy speech. On our corpora, a good value of n is 3.
The static labelling consists in comparing the means and standard deviations of S by segment to those thresholds.

Temporal coordination : We use the following rules for each segment : an isolated speech segment is classified "noise", a short noise segment between two speech segments is classified "speech" if its duration is less than 80 ms (maximum of a plosive silence). Last, exceptions are taken into account in order to manage some particular cases (very low SNR, impulsive noise near a speech zone). At the end of this step, we own the final labelling of the segments, and so, the position of speech and noise zones in the signal.

2.3.2. Mean noise vector estimation

The best way to evaluate the noise estimation $|\hat{B}(\omega)|$ is to use the longest noise zone available in the temporal domain. This zone is divided in frames of 256 samples, with a 128 samples overlapping. Once in the spectral domain, each frame will give a short term power spectral density vector. $|\hat{B}(\omega)|$ is computed as the mean of these vectors.

2.4. Noise compensation

We have implemented the Linear Spectral Subtraction algorithm, this method consists in subtracting the estimate of the noise power spectral density to the power spectral density of each frame of noisy speech. One of the risks of this method is to operate a too brute subtraction in the low SNR zones, where speech

signal can be confused with noise, so when SNR is too low (less than 5 dB), we use a weight coefficient (proportional to the SNR) to attenuate the effects of the subtraction. Effectively, our priority is to privilege the speech signal integrity on the noise compensation efficiency.

3. ROBUST NETWORK ARCHITECTURES

3.1. Introduction

To take more variable conditions into account, it is current to increase the number of HMM parameters, the more classical method consists in using multi-gaussian distributions [3].
We propose a more supervised increasing of the number of the model parameters, by specific network modelling. The idea is to provide more variability to the system to be trained, and to support this variability with this greater number of parameters.
[9] and [4] proposed solutions aiming to model separately speech and noise in a HMM, our approach is rather different : we introduce two robust HMM architectures :

- Multi-HMMs are built by gathering several HMMs. A specific database is collected from each noise context to train a primary HMM; the noise contexts correspond to different kinds of telephonic networks (analog, digital, cellular phone). A multi-HMM is obtained by gathering the initial and final states of each primary HMM.
- Multi-transitions models are designed to improve the number of probability density functions of the model, by improving the number of transitions between each state. As previously, we use various noisy speech corpora to initialize each law corresponding to a noise context. For instance, one law is initialized using telephonic data, another one is initialized with cellular phone data, and a last law is initialized with both kind of data. Then, the whole system is trained using all available corpora.

3.2. Multi-HMMs

The idea of this method is to use different parallel networks for each unit to recognize. A classical exemple consists in modelling separately male and female speakers. Both model are trained separately on two partitions of the corpus (men and women), then gathered by a common beginning and ending state.
In our application, multi-HMMs are built separately and each one is trained on a corpus relative to a noise context. The different noise contexts correspond to different kinds of telephonic networks (analogic, digital, cellular phone).
Then, the first and last state of each HMM is connected to a common beginning and ending state, as presented in figure 1.

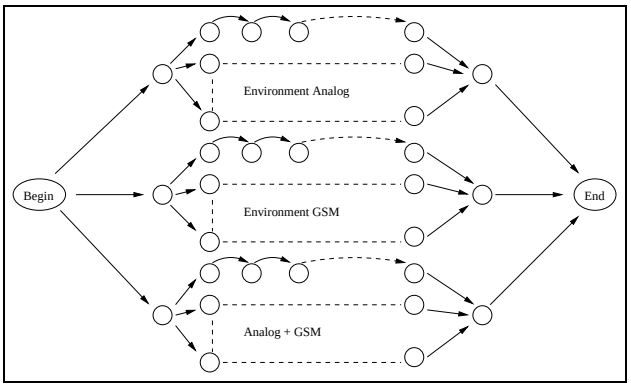

Figure 1: Exemple of multi-HMMs.

Using n HMMs, this method multiply by n the number of states and transitions of the original model.

3.3. Multi-transitions

Our objective is to improve the number of probability density functions of the model. In our HMM structure, Gaussian laws are associated with the transitions, so we multiply the number of transitions between each state to improve the number of laws (figure 2). So, multi-transition models use a partition of the data: various noisy speech corpora are used to initialize each law corresponding to a noise context. Then, the whole system is trained using all available data.

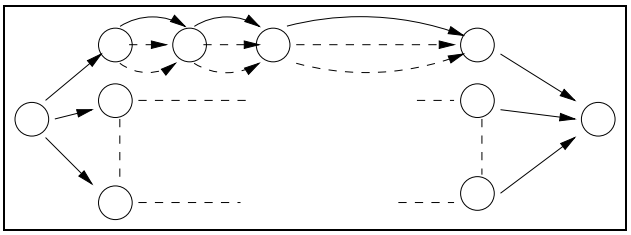

Figure 2: Exemple of multi-transition.

Using m probability density functions, this method multiply by m the number of *pdf* of the original model.

3.4. Multi-HMMs of multi-transition

These two methods can be combined in multi-models of multi-transitions: the number of parameters is significally augmented, the learning corpora include more variability, so that the system is drastically more robust. With n HMMs, each using m transisitions between each state, we multiply by n the number of states and by $n \times m$ the number of transitions of the original model.

4. EXPERIMENTATION

All these methods have been tested on two CNET corpora of 16 words pronounced by one hundred speakers : a telephonic and a cellular phone corpus, with a segmental pseudo-diphone HMM recognition system using continuous densities laws.
Only one acoustic vector is used for each segment : this vector corresponds to the spectral analysis of the central frame of the segment. We use 8 Mel-Frequency Cepstral Coeffients and 8 derivatives obtained by regression on the neighbouring frames, energy of the frame and its derivative, and length of the segment.
The average error rate of our system is 13.2 % when tested in mismatched conditions. Noise compensation presents an average error rate of 9.6 %. Figure 3 shows error rates evolution when using robust architectures modelling; each set of columns presents results obtained on a telephonic environment.
First and second columns of both sets present, for comparison, results obtained in matching conditions and with a basic mixed training (all available data of both environments used to train an ordinary HMM). Last 3 columns present results of the robust architectures: multi-HMMs reduce the error rate of 64 % (col. 3); multi-transitions reduce this rate of 70 % (col. 4); multi-HMM of multi-transition reduce the error rate of 77 % (col. 5).

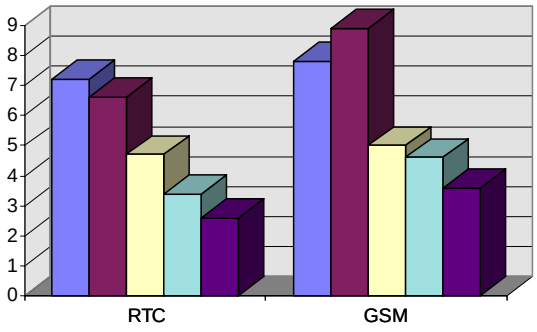

Figure 3: Results of experimentations.

5. CONCLUSION

Robust architectures are drastically more efficient, but the use of this method depends on the existence of learning corpora corresponding to various noise contexts. When it is possible to record the same speech corpora in various noise contexts, the robust architectures achieve very good performances for building a unique model.
When impossible, noise compensation or channel equalization are the only affordable solutions.
Evolution of the method include a combination of network modelling and channel effect compensation by cepstral subtraction [7].

References

[1] R. André-Obrecht. A new statistical approach for automatic segmentation of continuous speech signals. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36:26–40, january 1988.

[2] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 27(2):113–120, april 1979.

[3] L. Fissore, P. Laface, and P. Ruscitti. HMM modeling for speaker independant voice dialing in car environment. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 249–252, San Francisco, California, april 1992. ICASSP'92.

[4] M. Gales and S. Young. An improved approach to the Hidden Markov Model decomposition of speech and noise. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 233–236, San Francisco, California, april 1992. ICASSP'92.

[5] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon. An improved endpoint detector for isolated word recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 29:777–785, august 1981.

[6] D. Mansour and B. Juang. A family of distorsion measures based upon projection operation for robust speech recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37:1659–1671, 1989.

[7] C. Mokbel, D. Jouvet, and J. Monné. Blind equalization using adaptative filtering for improving speech recognition over telephone. In *Proc. Eur. Conf. on Speech Communication and Technology*, pages 1987–1990, Madrid, Spain, september 1995. EUROSPEECH'95.

[8] J-B. Puel and R. André-Obrecht. Robust signal preprocessing for HMM speech recognition in adverse conditions. In *Proc. Int. Conf. on Spoken Language Processing*, pages 259–262, Yokohama, Japan, september 1994. ICSLP'94.

[9] A. Varga and R. Moore. Hidden markov model decomposition of speech and noise. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 845–848, Albuquerque, New Mexico, april 1990. ICASSP'90.