CELLULAR PHONE SPEECH RECOGNITION : NEURAL NETS PREPROCESSING vs. ROBUST HMM ARCHITECTURES

Jean-Baptiste PUEL
IRIT - Université Paul Sabatier
Toulouse - FRANCE

ABSTRACT

In this paper, we present and compare two methods contributing to the robustness of automatic speech recognition systems in adverse conditions. The first method consists in computing new acoustic parameters corresponding to the identification of a telephonic network using a neural net. The second one consists in building robust HMM architectures to include and manage more variability in the learning corpora.

INTRODUCTION

Speech recognition in adverse conditions require specific processing for dealing with additive noise and for channel effect compensation. A classical approach consists in preprocessing data before the recognition task, in order to make these data as close as possible from the system references : the data used to train the recogniser.
This method considers that the modification suffered by the acoustical parameters can be modelled by

- addition of noise in the spectral domain [Boll79]
- convolutive filtering in the temporal domain (channel effect) [Atal74]

Depending on the noise kind, good results can be obtained by spectral or cepstral subtraction methods, but in some cases, this solution is not adequate.
For instance, in the case of interactive vocal servers, adaptation to telephonic lines and to cellular phone is a much complex problem and the basic methods we cite do not achieve good performances. In that context, we propose two kinds of methods to improve the performances of automatic speech recognition systems used on various communication channels :

- a neural nets method aiming to identify the noise context and add new information to the acoustic coefficients used by an HMM recognition system,
- robust HMM architectures (multi-HMMs and multi-transitions) that take into account more variability in the data used to train the systems.

NEURAL NETS PREPROCESSING

Introduction

A general problem for automatic speech recognition on unknown communication channels is the identification of channel : when it is possible to identify on which kind of telephonic network a

communication has been established, it becomes possible to adapt the parameters of the speech recognition system.
We present a novel neural net method aiming to identify the communication channel by the identification of the channel noise. The results of this identification are used in a HMM recognition system.

Neural Net identification of noise context

The idea of our method is to train a neural network with data recorded in various noise environments. The input of the system is a vector of acoustic coefficients corresponding to a signal frame and computed with cepstral analysis, the output is the environment identified by the network.
Experiences have been made using the following steps :

- collecting an important database of acoustic vectors (about 7,000 samples) by automatic alignment of an HMM on pseudo-diphones extracted from words pronounced by one hundred speakers, on telephonic lines and cellular phones.
- balance the data so that each sound (vowels, consonants, silence frames) be represented in same number in each environment.
- training of a neural network to identify the noise context, using various configurations of entries, number of neurons, size of the hidden layer.

The neural networks we used are multi-layers perceptrons with backprop learning algorithm.
Best results have been obtained for the following configuration :

- vectors of 8 Mel Frequency Cepstral Coefficients and energy of the frame (derivatives bring no results to this kind of neural nets, unable to manage time information)
- 9 neurons in the input layer, 5 in the hidden layer and 2 neurons in the output layer

For an isolated frame, the average recognition rate of the correct noise environment is 70 %, but using all the frames of an utterance, the average recognition rate of a word is over 80 %.
It is important to note that the best recognition performances are obtained on speech frames, and not on noisy silence frames : it seems that most of the chanel effect information in the cepstral domain is carried by speech frames vectors. A phenomenon of this kind was yet described in [Mokbel et al.95].

Experimentation

Using the neural network trained to identify the telephone line noise (analogic or cellular phone), we process the learning data of an HMM speech recognition system, in order to add two new coefficients to each vector, corresponding to the results of the output layer of the neural net.
Two CNET speech corpora are used, recorded on both telephonic networks, each one include 16 words pronounced by one hundred speakers.
The recognition system trained with these data presents a reduction of the error rate of about 25 % (6 % to 4.5 %) : the new coefficients bring pertinent information to the system.

Originality of this work resides in the fact that the neural network is not used to modify acoustic space of the data (like in [Tamura89],[Sorensen91] or [Ohkura et al.91]), but to classify the data and give the results to an HMM.
Results obtained by this hybrid system are promising, so we are currently testing other kinds of neural network on the same task.

ROBUST HMM ARCHITECTURES

Introduction

Speaker independant small vocabularies speech recognition seems to be a solved problem using Hidden Markov Models. Such systems offer realistic solutions for creating interactive vocal servers, but the lack of robustness to environment changes is their principal weakness [Gong95].
In order to obtain good recognition performances in adverse conditions, the learning data must be as close as possible to the real using conditions data [Juang91]. Generally, it is impossible to know in advance in which noise conditions the system will be used.
So, we propose a set of robust HMM architectures including multi-HMMs and multi-transitions modelling to enhance recognition performances of systems designed to be used in different noise conditions. We present the results obtained by these methods on CNET speech corpora recorded with telephonic and cellular phone noise conditions (European GSM system).

Robust HMM architectures

The characteristics of telephonic noise not only depend on the signal to noise ratio : the acoustic chanel acts as a convolutive filter on the speech signal, the cellular phone coders introduce important degradations including impulsive noise.
The method we propose to enhance systems robustness consists in increasing the number of parameters used to describe the vocabulary of the application, by specific network modelling. The idea is to provide more variablity to the system to be trained, and to support this variability with this greater number of parameters.
[Varga et al.90] and [Gales et al.92] proposed solutions aiming to model separately speech and noise in a HMM, our approach is rather different : we introduce multi-HMMs corresponding to each noise contexts, and multi-transitions on a single model to contain the description of an utterance pronounced in various noise contexts.
Multi-HMMs are built separately and each one is trained on a corpus relative to a noise context. In our application, the different noise contexts correspond to different kinds of telephonic networks (analogic, digital, cellular phone), used for training the HMMs. Then, the first and last state of each HMM is connected to a common beginning and ending state, as presented in the next figure :

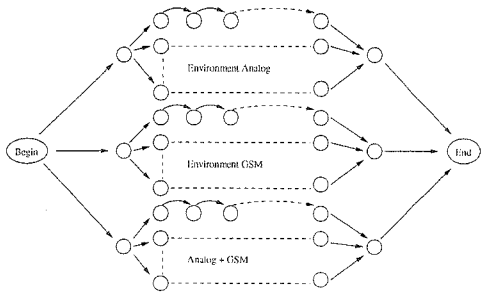

Figure 1: Exemple of multi-HMMs.

Multi-transitions systems 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 data.

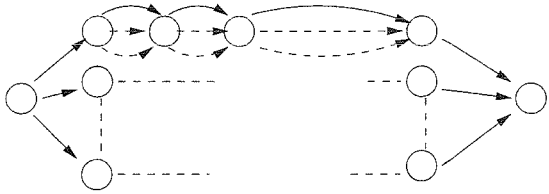

Figure 2: Exemple of multi-transition.

The two methods can be combined in multi-HMMs of multi-transitions : the number of parameters is significally augmented, the learning corpora include more variability, so that the system is drastically more robust.

Experimentation

All these methods were tested on two CNET corpora of 16 words pronounced by one hundred speakers : a telephonic and a cellular phone corpus, with a pseudo-diphone HMM recognition system using continuous densities laws.
The multi-HMMs reduce the error rate from 13.2 % to 4.8 % (64 % amelioration).
The multi-transitions reduce this rate from 13.2 % to 4.0 %. (70 % amelioration).
Multi-HMMs of multi-transitions from 13.2 % to 3.1 % (77 % amelioration).
Evolution of the method include a combination of network modelling and chanel effect compensation by cepstral subtraction [Mokbel et al.94].

CONCLUSION

The robust HMM methods offer very better performances than the neural nets methods, due to a very larger number of parameters used in the modellisation.
On the other hand, the 25 % improvment of the neural nets method can be compared to classical noise compensation methods, *without changes on the signal*, only a new parameter is added to the acoustic vectors.
Next improvment of the methods include a collaboration of the two algorithms : the neural net identification of the telephonic network will guide the multi-HMMs to choose the best HMM, or the multi-transitions system to use the correct transition.

References

[Atal74] ATAL (B). – Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of The Acoustical Society of America*, no55, 1974, pp. 1304–1312.

[Boll79] BOLL (S). – Suppression of acoustic noise in speech using spectral subtraction. *IEEE TRANS. on ASSP*, vol. 27 (2), april 1979, pp. 113–120.

[Gales et al.92] GALES (M) and YOUNG (S). – An improved approach to the hidden markov model decomposition of speech and noise. *In : Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. ICASSP'92, pp. 233–236. – San Francisco, California, april 1992.

[Gong95] GONG (Y). – Speech recognition in adverse environments : a survey. *Speech Comunication*, vol. 16, 1995, pp. 261–291.

[Juang91] JUANG (B). – Speech recognition in adverse environments. *Computer Speech and Language*, vol. 5, 1991, pp. 275–294.

[Mokbel et al.94] MOKBEL (C), PACHÈS-LÉAL (P), JOUVET (D) and MONNÉ (J). – Compensation of telephone line effects for robust speech recognition. *In : Proc. Int. Conf. on Spoken Language Processing*. ICSLP'94, pp. 987–990. – Yokohama, Japan, september 1994.

[Mokbel et al.95] MOKBEL (C), JOUVET (D) and MONNÉ (J). – Blind equalization using adaptative filtering for improving speech recognition over telephone. *In : Proc. Eur. Conf. on Speech Communication and Technology*. EUROSPEECH'95, pp. 1987–1990. – Madrid, Spain, september 1995.

[Ohkura et al.91] OHKURA (K) and SUGIYAMA (M). – Speech recognition in a noisy environment using a noise reduction neural network and a codebook mapping technique. *In : Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. ICASSP'91, pp. 929–932. – Toronto, Canada, may 1991.

[Sorensen91] SORENSEN (H). – A cepstral noise reduction multi-layer neural network. *In : Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. ICASSP'91, pp. 933–936. – Toronto, Canada, may 1991.

[Tamura89] TAMURA (S). – An analysis of a noise reduction neural network. *In : Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. ICASSP'89, pp. 2001–2003. – Glasgow, UK, may 1989.

[Varga et al.90] VARGA (A) and MOORE (R). – Hidden markov model decomposition of speech and noise. *In : Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. ICASSP'90, pp. 845–848. – Albuquerque, New Mexico, april 1990.