SOLUTIONS FOR ROBUST RECOGNITION OVER THE GSM CELLULAR NETWORK

Lamia Karray, Abdellatif Ben Jelloun and Chafic Mokbel

FT.CNET/DIH/RCP
2 av. Pierre Marzin
22307 Lannion cedex, France
E-mail: lamia.karray@cnet.francetelecom.fr

ABSTRACT

This paper deals with automatic speech recognition robustness for noisy wireless communications. We propose several solutions to improve speech recognition over the cellular network. Two architectures are derived for the recognizer. They are based on Hidden Markov Models (HMMs) adapted to adverse noise conditions. Then two more specific solutions aiming to alleviate GSM cellular network defects (holes and impulsive noise) are developed. Holes are detected and rejected. Impulsive noises are modeled using mixture density HMMs and a maximum likelihood criterion. These solutions allow a noticeable recognition error reduction. The last one seems to be promising.

1. INTRODUCTION

Speech recognition systems operating in a practical environment may have to deal with a wide variety of disturbances. Besides, the rapid development of cellular networks offers new opportunities for the application of speech recognition. However, this involves adverse environments of use such as running cars, public places, etc. In such conditions, noise is very disturbing and robustness becomes an important requirement for successful use of speech recognizers.

Several solutions for robustness improvement were considered. For example, spectral subtraction to improve robustness to additive noise [1, 2]; or cepstral normalization for convolutive distortions [3, 4]. Modeling speech and noise separately was also proposed [5, 6].

In this paper, we consider recognition over the GSM network in adverse conditions. We first describe, in section 2, the problems of speech distortions in a database collected over such network. Then, we propose two different solutions based on robust HMM architecture described in section 3. In section 4, a detection algorithm of GSM holes is introduced and a rejection procedure is adopted. Section 5 describes an optimal model for GSM impulsive noises based on a Maximum likelihood criterion and using a combination between the initial HMM and a mixture Gaussian modeling the GSM noises.

2. COMMUNICATION OVER CELLULAR NETWORKS

Communications over cellular networks have sometimes very poor listening quality. In some environments, one can hear a clanging voice (impulsive noise) or a sudden disappearance of the signal (that we call "a GSM *hole*"). This temporary absence of signal, may affect the beginning or the end of a word, or even cut it into tow or more meaningless sounds.

We focus on the cellular Global System for Mobile (GSM) network. We consider a laboratory GSM database of 51 vocabulary words [2]. Several call environments are considered: indoors (26% of the considered calls), outdoors (22%), stopped cars (29%) and running cars (23%). In the GSM speech signal, ambient noises are frequent (especially in outdoor and running car calls), and the GSM transmission effects may be very disturbing.

The database is hand segmented and labeled. Therefore, different labels of noise and Out-Of-Vocabulary (OOV) words were added to the initial vocabulary words. This results in a database of 35995 segments including 64% of vocabulary words, 7% of OOV words and 29% of noise (16 % of ambient noises, 9 % of GSM channel distorsion and 4 % of remaining echoes).

3. ROBUST HMM ARCHITECTURES

It has been shown [7] that the best recognition results are obtained when the training data condition match the real condition data, where the system will be used. This is not feasible in practice for telephone applications since one can never know *a priori* in which environment a recognizer will be used.

Speech recognition performances decrease in noisy environments, typically for communications over mobile phones used outdoor or in running cars. However, significant improvements could be achieved when the same environment is considered for training and recognition.

In order to overcome this problem, various algorithms have been proposed [8]. Since the condition in which the

system will be used is not known *a priori*, we propose a system able to choose the appropriate one between different conditions.

The idea is to train the system using a greater number of parameters, in order to provide precise modeling able to take more variability into account. This can be achieved by means of multi-models or multi-transition modeling [8].

3.1. Multi-HMMs

The idea consists in combining several HMMs trained separately on different databases, related to different call environments. The global HMM has common beginning and ending states, and can choose between the different models.

3.2. Multi-transition model

In this case, different output pdfs (probability density functions) are associated to each state. Each pdf is initialized using the corresponding pdf in a HMM trained in a given condition. During the decoding phase, Viterbi algorithm chooses the most likely pdf. This is practically achieved using different transitions between two states, a condition pdf is associated to each transition.

The whole model is then retrained on the global data. In the recognition phase, the system is free to choose the most suitable transition.

3.3. Experimentation

To build the different models, we use the CNET HMM-based system (PHIL90) [9].

In our case, we distinguish two conditions :

- quiet environment including indoor or stopped car calls,
- noisy environment including outdoor or running car calls.

Data collected in these environments are used to build two 30-state HMMs. These HMMs are then used to initialize a bi-model HMM and a bi-transition one. Training and recognition phases are then performed using data collected in various conditions. Figure 1 summarizes the results.

We notice an improvement with bi-transition model.

More improvement could be achieved by the combination of the two architectures above. This yields a multi-HMMs of multi-transitions models. With this architecture, the number of parameters is increased, the learning data includes more variability, which should increase the robustness of the system. However, we did not test this system because the high number of parameters needs much more data than we have to achieve a correct training.

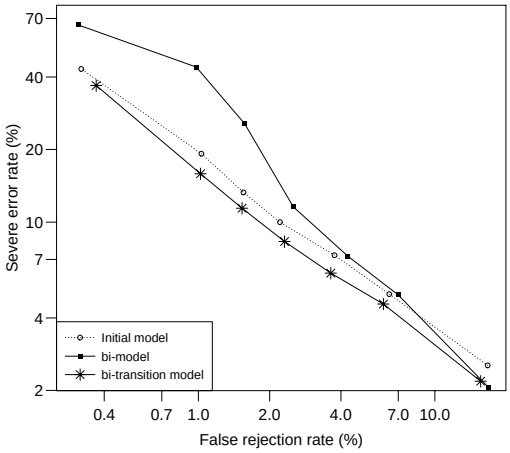

Figure 1: Robust architectures results. Severe errors (i.e., substitution and false acceptance errors) with the initial model, the bi-model and the bi-transition one are plotted. The different points of the plot are obtained by varying the weight of the garbage model probabilities.

4. GSM HOLE REJECTION

In presence of holes or clanging noise, recognition becomes difficult and error rates increase significantly, especially substitution errors and false rejection.

In this section we consider the GSM holes. To avoid this problem, we developed an algorithm allowing to detect the "holes" through the GSM communications in order to favour their rejection during the decoding phase.

Holes detection algorithm is based on some statistical properties of those segments of the signal. Amplitude, variance and segment length are taken into account. It has been noticed that the corresponding segments have low amplitude (about 5) and variance (about 3.5) compared to the segments of silence (see figure 2).

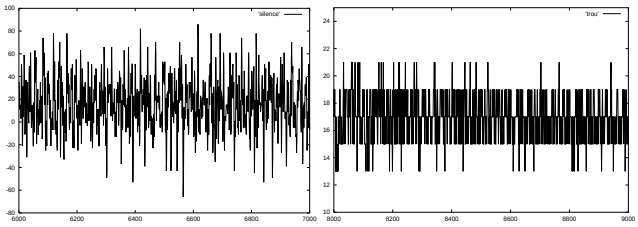

Figure 2: Difference between silence (left) and a GSM hole (right).

When a hole is detected, the associated word has less chance to be correctly recognized. Thus, we favour its rejection by emphasizing the garbage model probability in the unigram used in the recognition. Hence, the substitution errors and false rejection rates are decreased as shown in figure 3.

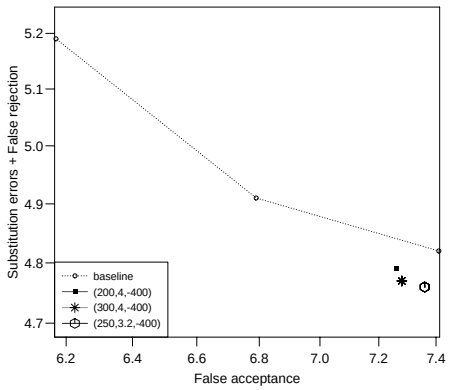

Figure 3: GSM hole rejection. Results are obtained using the initial model without (*"baseline"* curve) and with the hole rejection procedure. The different points of the *"baseline"* plot are obtained by varying the weight of the garbage model probabilities. The other points result from various sets of *(hole lenght,variance,rejection weight).*

5. MULTI-GAUSSIAN DISTRIBUTION MODELING OF GSM NOISE

In the training GSM database, noise segments were hand labeled so that they could be used to obtain a rejection (garbage) model. We used Gaussian HMM models for vocabulary words as well as for noise (rejection) tokens. In order to avoid the very disturbing GSM noise, we will try to improve its modeling. Using *a priori* knowledge of the GSM noise, this method aims to give a more precise model of the vocabulary words. Two HMMs are combined: one for the whole vocabulary words and the other for GSM noise. For GSM noise modeling, we use a mixture Gaussian model allowing to make a more precise estimation of the observation law. Then, in a given call, the observation law is optimized using a maximum likelihood criterion. This means that the Viterbi algorithm has to choose between two observation laws: the one obtained by a mixture Gaussian noise modeling and the initial Gaussians of the HMM model.

Recall that, for a given HMM λ, an observation sequence $O_1, \ldots, O_t$ and a set of $q_1, \ldots, q_t$ possible paths to the state s_i, we define the corresponding optimal path as follows:

$$\delta_t(i) = \max_{q_1,\ldots,q_{t-1}} P(q_1, ..., q_{t-1}, q_t = s_i; O_1, ..., O_t/\lambda)$$

The optimal path is obtained recursively using the formula :

$$\delta_{t+1}(j) = \max_{1 \le i \le N} \{\delta_t(i) a_{ij}\} b_{ij}(O_{t+1})$$

where N is the state number in the HMM, $a_{ij} = P(q_{t+1} = s_j / q_t = s_i)$ and $b_{ij}(O_{t+1}) = p(O_t / q_t = s_j, q_{t-1} = s_i)$.

In this case, the pdf b_{ij} results from a combination of two possibilities: $b_{ij}^{initial}$ given by the initial Gaussians of the HMM trained on the global database words, and B_{ij} given by a Gaussian mixture HMM optimized on GSM noise tokens. As a combination, we chose the following:

$$b_{ij}(O_t) = \max\{(1-\beta) b_{ij}^{initial}(O_t), \beta B_{ij}(O_t)\}$$

where β is a weight aiming to favour one of the combined densities.

Let us now detail the GSM noise modeling. In practice, GSM noise is supposed to be independent of the clean speech signal. To model it, we consider the particular case of one state HMM with a Gaussian mixture density. The observation law at time t is then:

$$B(O_t) = \sum_{k=1}^{N_g} c_k N(O_t, \mu_k, \Sigma_k) = \sum_{k=1}^{N_g} c_k N_k(O_t)$$

where c_k is the weight of the k^{th} component in the mixture of N_g Gaussian functions. Gaussian mixture parameters are optimized in the training phase using only GSM noise hand labeled tokens. The optimization is performed by means of the iterative EM algorithm.

This algorithm needs an initial parameter vector. A starting point is obtained by means of vector quantization of the acoustic sub-space of the training parameter vectors. Vector quantization aims to split the whole space of parameter vectors into a given number of classes (which is the number of Gaussian functions in the mixture density). Vector quantization is performed using the LBG (Linde, Buzo, Gray) algorithm [10].

The initialization associates each Gaussian to a given class. Hence, the initial mean vector of the Gaussian number k is the centroid of the k^{th} class, and the covariance matrix is computed using the observation vector of that same k^{th} class.

As for the weights c_k, we chose a uniform repartition as an initial weight vector: $c_k = \frac{1}{N_g}$ where N_g is the number of Gaussian functions in the mixture.

Notice that since the GSM noise model involves only one state, there is only one pdf $B(O_t)$ for each observation O_t. It is estimated as the best Gaussian function in the mixture: $B(O_t) = \max_k \{c_k N_k(O_t)\}$.

Hence, the training phase consists in building this GSM noise model. Then, in order to improve the recognition performances, this model will be combined with the initial 30-state HMM trained on the whole vocabulary, including vocabulary words and noises. The one with the highest likelihood is selected by the Viterbi algorithm during the decoding phase.

The evaluation of this method is performed using the classical test procedure giving recognition error rates. Sev-

eral tests were conducted varying the weight β and the Gaussian functions number in the mixture density (N_g). Results are summarized in figure 4.

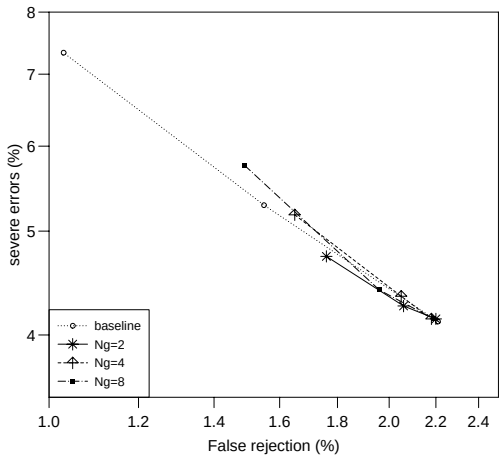

Figure 4: GSM noise modeling. Results are obtained using the initial model combined with the Gaussian mixture associated to the GSM noise model. *"baseline"* curve corresponds to the initial HHM model (without mixture). The other curves are obtained by varying the number of Gaussians in the mixture (Ng). The different points on each plot are associated to a weight β of the mixture in the combination.

We notice that the best improvement is achieved with two Gaussian functions in the mixture. For instance, for 2.1% of false rejection errors, we achieve 3% decrease of substitution error rates and 8% decrease of false acceptance error rates. However, we believe that more improvement could be obtained using this technique with more GSM noise tokens in the mixture training data.

6. CONCLUSION

In order to improve robustness to noise in automatic speech recognition over cellular networks, several techniques were described in this paper.

First, we introduced two architectures taking the environment call into account: multi-models and multi-transition models. Bi-transition models was shown to outperform both of the bi-model architecture and the initial one. For instance, it achieves 18% of severe error rate reduction for a given false rejection rate (1%).

Then, we introduced two techniques more specifically adapted to the GSM "holes" and impulsive GSM noise. Both of them improved the recognition rates. In the second one, the use of a Gaussian mixture provides a precise modeling of the GSM noise without any *a priori* knowledge of the decoding procedure within the cellular network transmission phase. This technique seems to be promising. Therefore, further developments based on this technique are being conducted in our team.

7. REFERENCES

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

[2] C. Mokbel, L. Mauuary, L. Karray, D. Jouvet, J. Monné, J. Simonin and K. Bartkova, "Towards Improving ASR Robustness for PSN & GSM Applications," to appear in Speech Communication.

[3] H. Hermansky, N. Morgan and H.G. Hirsch, "Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing," Proc. ICASSP'93, pp. 83-86, 1993.

[4] C. Mokbel, D. Jouvet and J. Monné, "Deconvolution of Telephone Line Effects for Speech Recognition," Speech Communication, Vol. 19, no. 3, pp. 185-196, September 1996.

[5] M.J.F. Gales and S.J. Young, "Robust Continuous Speech Recognition Using Parallel Model Combination", IEEE Trans. on Speech and Audio Processing, vol. 4, no. 5, Sept. 1996.

[6] S.V. Vaseghi and B.P. Miller, "Noise-Compensation Methods for Hidden Markov Model Speech Recognition in Adverse Environments," IEEE Trans. on Speech and Audio Processing, vol. 5, no. 1, January 1997.

[7] B. Juang, "Speech Recognition in Adverse Environments," Computer Speech and Language, no. 5, pp. 275-294, 1991.

[8] J.B. Puel and R. André-Obrecht, "Cellular Phone Speech Recognition: Noise Compensation vs. Robust Architectures," Proc. Eurospeech'97, pp. 1151-1154, Rhodes, Greece, 1997.

[9] C. Sorin, D. Jouvet, C. Gagnoulet, D. Dubois, D. Sadek, and M. Toularhoat, "Operational and Experimental French Telecommunication Services Using CNET Speech Recognition and Text-To-Speech Synthesis," Speech Communication, Vol. 17 (3-4), pp. 273-286, 1995.

[10] Y. Linde, A. Buzo and R.M. Gray, "An Algorithm for Vector Quantization Design," IEEE Trans. on Communications, vol. COM-28, no. 1, pp. 84-95, 1980.