NOISE MODEL SELECTION FOR ROBUST SPEECH RECOGNITION

L. Docío-Fernández and C. García-Mateo,

E.T.S.I. Telecomunicación
Departamento de Tecnologías de las Comunicaciones
Universidad de Vigo, 36200 Vigo, SPAIN.
E-mail: ldocio@tsc.uvigo.es, carmen@tsc.uvigo.es

ABSTRACT

This paper addresses the problem of mismatch between training and testing conditions in a HMM-based speech recognizer. Parallel Model Combination (PMC) has demonstrated to be an efficient technique for reducing the effects of additive noise. In order to apply this technique, a noise HMM must be trained at the recognition phase. Approaches that estimate the noise model based on the Expectation-Maximization (EM) or Baum-Welch algorithms are widely used. In these methods the recorded environmental noise data are used, and their major drawback is that they need a long sequence of noise data to estimate properly the model parameters. In some real life applications the amount of noise data can be too small, so from a practical point of view, the needed amount of noise is a critical parameter which should be as short as possible. We propose a novel method for obtaining a more reliable noise model than training it from scratch by using a short noise sequence.

1. INTRODUCTION

The accuracy of automatic speech recognition systems (ASR systems) rapidly degrades when there is a ***mismatch*** between the training and testing conditions. It has been demonstrated that ASR systems can perform very poorly when they are tested using a different type of acoustical environment from the one with which they were trained. Therefore, the requirement for **Robust Speech Recognition Systems** is becoming increasingly important as they are applied to practical applications. Applications such as speech recognition over telephone, in cars, on a factory floor or outdoors demand a great degree of environmental robustness.

The goal of Robust Speech Recognition is to minimize the effects of such a mismatch, so as to obtain a recognition accuracy as close as possible to that obtained under matched conditions. A wide variety of schemes for dealing with the problem of robust speech recognition has been proposed. In this paper we will focus on the Parallel Model Combination (PMC) scheme [1], which has demonstrated to be an efficient technique for reducing the effects of additive noise. Among all the PMC proposed approaches we will only consider the ***Log-Add Approximation*** because of its high computational efficiency.

In order to apply the PMC approaches a noise HMM (Hidden Markov Model) model must be trained "*on-line*" at the recognition phase before the speech model compensation process. This requires enough noise frames to estimate a reliable noise HMM. In this paper we present a novel technique for selecting, in real-time, a ***simple noise model*** from ***a general noise model*** rather than training it from scratch.

2. THE LOG-ADD PMC TECHNIQUE

PMC is a model-based noise compensation scheme for robust speech recognition. The aim of PMC is to alter the parameters of a set of HMM-based acoustic models estimated in a clean environment, so that they reflect the speech spoken in the current operation environment. The technique assumes that a clean speech model is available and a simple additive noise model can be estimated *on-line* to characterize the actual noise conditions.

The first stage of any robust recognition technique is to define some model of distortion or *mismatch function*, that describes the effects of the noise on the *clean* speech. The basic and most common model is that the original speech signal may be distorted by both additive and convolutional noise. Thus, the basic assumption behind PMC is that the speech and ambient noise are additive in the linear spectral domain, and the speech and channel distortion are multiplicative in that domain. Besides, the convolutional noise is assumed to be constant over time. Another required and important assumption at the PMC framework is that the frame/state alignment, used to generate the speech models from the clean speech data, is not altered by the addition of noise. With this, the *mismatch function* can be represented in the log spectral domain as

$$O_i^l(t) = H_i^l(t) + \log\left(g\,\exp\left(S_i^l(t)\right) + \exp\left(N_i^l(t)\right)\right)$$

For the purposes of the present paper the channel distortion has not been considered and only the effect of additive noise has been investigated. So, the mismatch function may be simplified to

$$O_i^l(t) = \log\left(g\,\exp\left(S_i^l(t)\right) + \exp\left(N_i^l(t)\right)\right)$$

Among the various PMC approximations, we only consider the **Log-Add Approximation**. It is the simplest and most computationally efficient PMC implementation. The model variances are assumed to be small ignoring so their effects on the estimates. This allows us to obtain a simplified and appropriate mismatch function for the static and dynamic coefficients that adapts only the HMM means while the variances are kept the same as those of the clean models [1].

It is important to note here that to lower both the computational overhead behind PMC technique and the recognition time[1], a simple noise HMM should be used. Thus, instead of using a more complex model, a single-state single-component CDHMM will be used throughout. Of course, this simple model does not represent properly the statistics of "complex" or non-stationary noise data such as telephone noise. However, for the comparison purposes, the simplest HMM-based noise model suffices. This greatly reduces the number of components and the model complexity in the robust recognition system.

3. THE NOISE MODEL SELECTION

As previously stated, to apply the PMC schemes a noise model estimate is needed to characterize the noise conditions of the current usage environment. The noise HMM is trained at the recognition stage prior to the model compensation process. This implies having enough environmental noise data to get a reliable noise HMM estimate.

The necessary environmental noise data can be obtained from non-speech frames inside the utterance (speech pauses). Thus, to perform the noise data selection in a practical recognition system some appropriate technique, such as a reliable Voice Activity Detector (VAD), is required to separate speech from stationary and non-stationary noises in the operation environment. By assuming a good speech frame classification accuracy of VAD, another important aspect in the performance of the noise compensation techniques is the available amount of noise data. In real-time recognition systems the length of noise data is a critical parameter which should be as short as possible. Normally, the noise model is determined over a finite length noise data. Clearly, a too short segment can not be used since the model estimate would be too unreliable.

The approaches commonly used to estimate the additive noise model in the compensation techniques obtain the test noise samples by recording the necessary background noise before the user starts to talk. The goal of the presented environmental compensation technique is to get a good recognition accuracy with a small amount of environment specific noise data and low computational cost. The proposed technique tries to avoid the on-line noise HMM training needed to implement the PMC schemes. The basic idea is to develop a technique of noise model selection that could be used in a real-time robust speech recognizer. The basic operations involved are:

1. By using a noise database, a single-state multiple-mixture general noise HMM model is trained *off-line*. This noise database must be representative of our possible operational scenarios.

2. At the recognition phase, we use the N first frames of the utterance to select one or more of the Gaussian mixtures as the actual noise model for the compensation process. So far, for this selection we have tried two different approaches:

   Technique A. With these N noise frames, we use the EM algorithm to update the mixture weights. The initial weights are then modulated to favor the ones which Gaussian is closer to the noise data. Therefore, the mixtures corresponding to the weights that have experimented the biggest growths will be selected to build up the actual noise HMM.

   Technique B. We decompose the multiple-mixture noise model into a set of single-state single-mixture noise models. Then, we apply a Viterbi (ML) decoder to the N first frames with a grammar that allows all the transitions among the noise models. The most frequently selected models will be used to obtain the actual noise HMM.

Figure 1 describes the above noise selection technique.

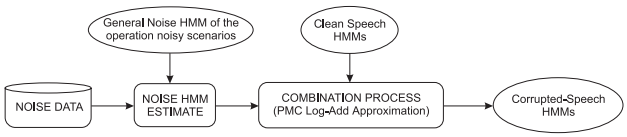

Figure 1: The proposed method for noise model selection.

Using the described method to obtain the noise model a set of Gaussians distributions, ranging from one to the number of components in the general noise HMM, are selected. So, to generate the simple noise HMM we need to apply some technique for merging or combining the individual Gaussian distributions. A variety of approaches can be used. The used approach in all the experiments was a simple average over all the chosen distributions. The mean of the noise HMM is calculated according to[2]

$$\hat{\mu} = \frac{1}{M} \sum_{m=1}^{M} \mu_m$$

where M is the number of selected Gaussians.

4. RECOGNITION EXPERIMENTS

Recognition experiments were conducted on a telephone speech database in order to evaluate the effectiveness of the proposed method.

4.1. The Recognition System

The recognition system is based on a set of 25 phone-like (PLU) Galician units, i.e., context independent (CI) units. In addition, six models to describe the silence and some noises, such as lipsmack, breath, and other background noises, were

[1] Inside the PMC framework a more complex noise model result in a recognition-time computational overhead due to increasing the number of components/states.

[2] To implement Log-Add PMC technique only the noise HMM mean is needed.

used. Each speech unit was modeled with a multiple mixture Gaussian Continuous Density HMM (CDHMM). The topology of the speech models was three left-to-right states. The noise source was modeled by a single-state single-Gaussian-mixture HMM. Diagonal covariance matrices were used throughout.

For training the HMMs we have used the HTK software [3]. And for the recognition experiments the HTK recognition system was used and appropriately extended in order to perform the Log-Add PMC Approximation.

The speech decoder uses a subword based Viterbi decoder constrained by a syntax consisting of silence followed optionally by some kind of background noises and by the task lexicon (one of the 655 possible *tokens*[3]).

4.2. Databases

The speech data used in the recognition experiments were extracted from a *multi-speaker Galician telephone* database called *"VOGATEL"* [2]. This database was collected over the telephone network from speakers calling from different regions of Galicia (Spain).

The input speech, sampled at 8KHz, was preprocessed using a 25ms hamming window and a 10 ms frame period. For each frame a set of 12 Mel-Frequency Cepstral Coefficients (MFCCs) were computed. The zeroth Cepstral coefficient was computed since it is needed in the PMC mapping process. Pre-emphasis (k=0.97) and liftering (L=22) were also used. The first and second order time derivatives, calculated using simple differences, were appended to the static parameters of each frame. This makes a 39-dimensional feature vector to represent each speech frame.

The training was done on a portion of the VOGATEL training corpus. We selected 2.035 "*clean speech*" files with a signal-to-noise ratio (SNR) above 30dB as training database. This database consisted of 651 male speakers and 2.742 different words extracted from phonetically balanced short utterances.

The testing database consisted also of a portion of the VOGATEL testing corpus. We selected 657 "*clean speech*" files with a SNR above 20dB as the "*clean test database*". This database consisted of 290 male speakers.

The work presented in this paper concentrates solely on data where the noise has been artificially added to each clean speech file. The noise sources considered were taken from the VOGATEL database. We have built a "*noise telephone database*" where a variety of telephone noises are available with a great temporal variability. The number of noise files is 1.059. This "*real-telephone noise database*" has been used to train *off-line* the **general noise HMM** that represents the possible operational noise scenarios.

In order to create the "*noisy test database*" each clean test file was corrupted by adding a noise file at 15dB SNR. To achieve this the original noise files have been properly scaled.

[3] A "token" is a single word or a short utterance.

4.3. Experimental Results

In this section we report various experiments using the previous databases and recognition system.

The PMC scheme used was the Log-Add approximation. Only the static parameters have been compensated since they are the most affected by the additive noise.

It is notoriously hard to obtain comparable results for a variety of "well-implemented" methods. However, it is worth examining some results in detail as they illustrate both their advantages and disadvantages. The aim of the experiments was to assess the proposed noise model selection techniques rather than to get the optimal performance.

First, to test the effect of the additive noise we have applied the baseline recognition system (no compensation) to the clean and noisy test files. Table 1 shows this baseline performance in terms of word accuracy. The column quoted "with C0" means that the zeroth cepstral coefficient (C0) was used in the recognition, and "without C0" means that C0 was dropped out. The addition of noise seriously degraded the performance reducing the recognition rate.

Condition	With C0	Without C0
Clean	84.93	85.54
Noisy (SNR = 15dB)	61.80	71.69

Table 1: Word accuracy rates (%) for the baseline system.

First, the two proposed noise model selection techniques (Technique A and Technique B) have been compared. The results are also shown in tables 2 and 3. First, both noise model selection methods were analyzed in terms of the number of chosen Gaussians. We have considered 1, 4, 8 or all the selected mixtures. The first aspect of the presented techniques assessed was how closely did the generated noise model match the trained model. It can be seen that better performance is obtained as the considered number of Gaussians increases, i.e., the model obtained to represent the noise is more accurate. When only one Gaussian is considered, the performance is very poor since the noise data is not being well represented. With regarding to the comparison between both techniques, we can say that by considering more than 1 mixture Technique B without C0 slightly outperforms Technique A. For 1 mixture Technique A gives better performance.

Second, we have analyzed how much noise data is needed to get an appropriate noise model in both the standard noise model estimation and the presented noise selection techniques. Both cases using and not using C0 have been considered. Tables 2 and 3 gives, respectively, the obtained word accuracy rate against the amount of noise frames used for the estimation. These results show that in some situations the new approaches outperforms the classical one, and in others a small degradation is observed. Comparing the obtained rates we can see that when the amount of noise data is less or equal to 40 frames (0.4 seconds) a "more appropriate" noise model can be obtained using one of the proposed methods, and for bigger amounts there is a slight degradation in performance compared to the standard method. For example, Technique B with "all the

selected Gaussians" and with 20 noise frames obtain the same performance (dropping out C0 in the recognition) that the standard method with 75 or 100 noise frames.

Model Set	Number of noise frames					
	10	**20**	**40**	**60**	**75**	**100**
Log-Add + EM	75.80	76.56	76.71	77.47	77.32	77.02
Log-Add + Tech A 1 Mixture	75.19	76.41	76.26	75.26	75.65	75.65
Log-Add + Tech B 1 Mixture	73.82	73.52	74.28	74.43	73.82	74.43
Log-Add + Tech A 4 Mixtures	76.10	76.71	76.41	76.71	76.26	76.56
Log-Add + Tech B 4 Mixtures	75.04	77.02	75.99	76.71	76.10	75.80
Log-Add + Tech A 8 Mixtures	76.41	77.47	77.47	77.63	77.32	76.86
Log-Add + Tech B 8 Mixtures	75.49	77.47	76.90	77.02	76.10	76.71
Log-Add + Tech A All Mixtures	76.26	77.63	77.02	77.47	77.02	77.32
Log-Add + Tech B All Mixtures	75.49	77.78	77.36	76.10	76.26	77.02

Table 2: Word accuracy rates (%) using C0 in the recognition for various techniques.

Model Set	Number of noise frames					
	10	**20**	**40**	**60**	**75**	**100**
Log-Add + EM	75.95	76.10	76.86	78.23	78.23	78.23
Log-Add + Tech A 1 Mixture	74.58	76.41	75.80	75.95	75.95	76.10
Log-Add + Tech B 1 Mixture	71.54	70.62	70.02	71.23	71.23	71.69
Log-Add + Tech A 4 Mixtures	76.10	77.17	76.56	77.47	76.71	76.86
Log-Add + Tech B 4 Mixtures	76.56	77.32	77.36	77.63	77.32	77.05
Log-Add + Tech A 8 Mixtures	76.26	77.17	77.47	77.47	77.63	77.17
Log-Add + Tech B 8 Mixtures	77.02	77.78	77.05	77.47	77.17	76.56
Log-Add + Tech A All Mixtures	76.41	77.17	76.86	77.32	77.32	76.56
Log-Add + Tech B All Mixtures	77.02	78.23	77.96	78.08	77.02	77.78

Table 3: Word accuracy rates (%) dropping C0 in the recognition for various techniques.

5. DISCUSSION

We have presented a novel method to obtain the noise model at the PMC framework. The performance in a complex-noise environment, the VOGATEL telephone noise, was examined. Recognition experiments confirmed that the proposed method improves recognition rates for noisy telephone speech. In addition, it requires a small amount of noise data and low computational cost for good performance.

The proposed techniques use a general noise model to create a simple and particular noise model. With these techniques it is possible to choose the number of Gaussians selected to generate the noise model. Better performance is obtained as the number of used Gaussians increases.

Experiments over a 15dB SNR noisy database were carried out. The performance with the amount of noise data was investigated. The experiments have show that few seconds of noise are enough to get the noise model. For small amounts of noise the proposed methods outperform the standard ones.

The efficiency and flexibility of the techniques and its adaptability to new situations make they suitable as the basis for a robust speech recognizer that is flexible to wide variations in conditions.

We are obtaining results with a noisiest condition. Specifically, the clean test database is corrupted at 10 and 5dB SNR. Other methods for combining the selected Gaussians are also being investigated.

6. ACKNOWLEDGEMENTS

This work has been partially supported by Spanish CICYT under the project TIC96-0964-C04-02.

7. REFERENCES

[1]. M. J. F. Gales, *"Model-Based Techniques for Noise Robust Speech recognition"*, PhD thesis, Cambridge University, 1995.

[2]. L. Villarrubia, P. Leon, L. Hernandez, J. M. Elvira, C. Nadeu, I. Esquerra, J. Hernando C. García-Mateo, L. Docío. *"VOCATEL and VOGATEL: Two Telephone Speech Databases of Spanish Minority Languages (Catalan and Galician)"*. Language Resources for European Minority Languages Workshop, Granada. (Spain) 26-31 May 1997.

[3]. S.J. Young et al., *HTK: Hidden Markov Model Toolkit V2.1*. Entropic Research Laboratories Inc., 1997.