Evaluation of Front-End Features and Noise Compensation Methods for Robust Mandarin Speech Recognition

R. Chengalvarayan

Lucent Speech Solutions
Lucent Technologies Inc.
2000 Lucent Lane, Naperville
Illinois 60566, USA
rathi@lucent.com

Abstract

This paper describes speaker-independent speech recognition experiments concerning acoustic front-end processing on a telephone database that was recorded in various dialect regions in China. In this paper, three different features based on human voice production, perception and auditory systems have been evaluated for Mandarin speech recognition. Experimental comparisons showed that auditory-filtered cepstral coefficients outperforms the other type of features. When speech recognizers are deployed in telephone services, they often encounter variable acoustic mismatches which significantly deteriorate their performance. Three different channel equalization techniques have been explored in this study to decrease this mismatch, hence improving the recognition accuracy. We present results with various noise compensation methods based on hierarchical cepstral mean subtaction and signal bias removal.

1. Introduction

In most state-of-the-art recognition systems, the speech signal is usually modeled by a set of hidden Markov models (HMMs) and the system performance deteriorates in mismatched training and testing conditions [4]. Robust systems perform equally well independent of the transmission channel characteristics, background noise, sound pick-up equipments, microphone response or modeling inaccuracies. In general, the acoustic mismatch can be reduced in several ways [5]. One way to reduce acoustic mismatches is to adjust speech features according to some models of the differences [7]. Another technique is to inject a fraction of the noise into the training data and retrain the system [1].

In this paper, several acoustic feature sets and different noise compensation methods have been evaluated based on their robustness to channel distortion for Mandarin telephone speech recogntion. Experimental comparisons using linear-predictive (LPCC), mel-frequency (MFCC), and auditory-filter (AFCC) based cepstral coefficients are given. The convergence property of the MSE training is investigated, and Mandrin connected-digit recognition results showed that the AFCC yields about 42% and 22% string error rate reductions in comparison with LPCC and MFCC systems. Further, we experiment with explicit channel compensation techniques such as cepstral mean subtractions (CMS) and hierarchical signal bias removal (HSBR). We demonstrate that the two-level CMS with AFCC delivers an additional string error rate reduction of 24%, 13%, and 8% when compared to that of the baseline, HSBR, and CMS systems.

2. Front-End Features

The structure of a typical continuous speech recognizer consists of a frontend feature extraction followed by a statistical pattern recognizer [6]. The feature vector, interface between these two, should ideally contain all the information of the speech signal relevant to subsequent classification and be insensitive to irrelevant variations due to changes in the acoustic environments [11]. Currently, there are three major approaches to the feature extraction based on modeling either human speech production or perception or auditory systems [2, 8, 10]. In this paper, the following three sets of acoustic features have been evaluated for robust Mandarin speech recognition.

- Linear prediction (LP) analysis is widely used in speech recognition for representing the short time spectral envelope information of speech. This analysis assumes the speech signal to follow an all-pole model and the importance of this method lies in its relative speed of computation. Most speech recognizers have traditionally utilized cepstral parameters derived from an LP analysis due to the advantages that LP provides in terms of generating a *smooth* spectrum, free of pitch harmonics, and its ability to model spectral peaks reasonably well.
- Mel-based cepstral parameters, on the other hand, take advantage of the perception properties of the human auditory system by sampling the spectrum at mel-scale intervals. A discrete cosine transform (DCT) is applied to the 24 overlapped mel-spaced triangular filter-bank log-energies and the first 12 cepstral coefficients are retained.
- An auditory feature extraction algorithm for robust speech recognition in adverse acoustic environments has recently been proposed based on the analysis of human auditory system [10]. It consists of 256-point FFT that generates a spectrum of 128 values at 8 KHz sampling rate. The magnitude spectrum is processed through a outer-middle-ear transfer function followed by Bark scale conversion. An auditory filter is then applied to smooth out the speech spectrum as in the cochlea. In the next step, the smoothed spectrum is passed through a logarithm or cubic-root function to stimulate the nonlinearith associated with auditory nerve discharge rates. Finally, a DCT is applied to convert the logarithmic spectrum to 12 cepstral coefficients.

The front-end processing for recognition is shown in Fig. 1. The overall system is a block processing model in which a 10

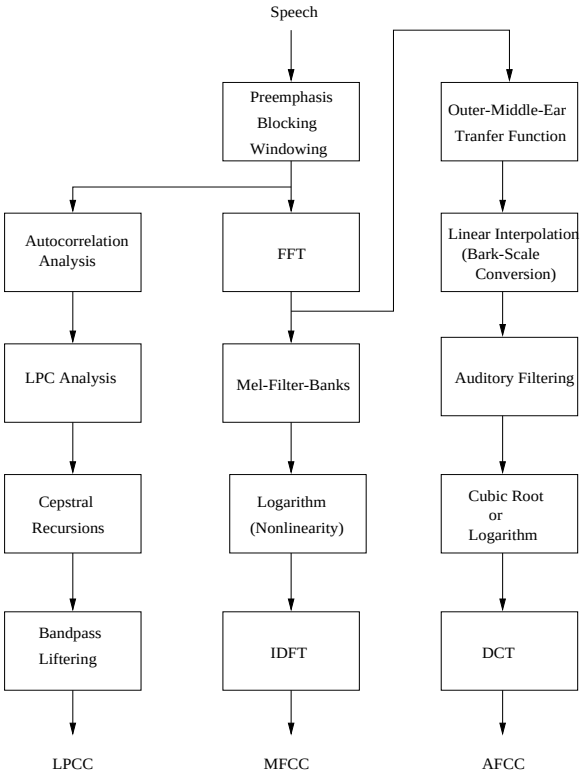

Figure 1: Block diagram of linear-predictive, Mel-frequency and auditory-filter based cepstrum generators.

msec frame is processed and a vector of static features is computed. Short-term energy by accumulating the power of the blocked speech samples before Hamming window is used as the energy term in all the features.

3. Noise Compensation Methods

It is well-known that a high-performance can not be maintained when there is a serious mismatch between training and testing conditions [4]. For telephone based speech recognition tasks, undesirable signal components can come from various sources, such as variations in telephone handsets, ambient noises and channel distortions. The speech distortion usually appears as a combination of various acoustic differences but the exact form of the distortion is often unknown and difficult to model [5]. To reduce these mismatches, some compensation is needed. This is usually performed in the feature space by modifying the feature vector of the testing signal closer to the trained models. In this paper, we investigate three different widely used noise compensation methods for robust telephone speech recogntion, namely, cepstral mean subtraction (CMS), two-level cepstral mean subtraction (2LCMS), and hierarchical signal bias removal (HSBR). All these methods improve the robustness of speech recognition systems by minimizing the effect of channel distortion on the measured signal input to the recognizer.

CMS is based on subtracting the long-term cepstral mean from each utterance. This idea is now widely used to remove mismatches due to time-invariant linear channel effects [3, 5, 7, 12]. In order to process the non-linear channel, a 2LCMS was proposed where separate channel compensation is performed for segments that are classified as speech and for segments classified as background [3]. The 2LCMS technique is implemented in several steps:

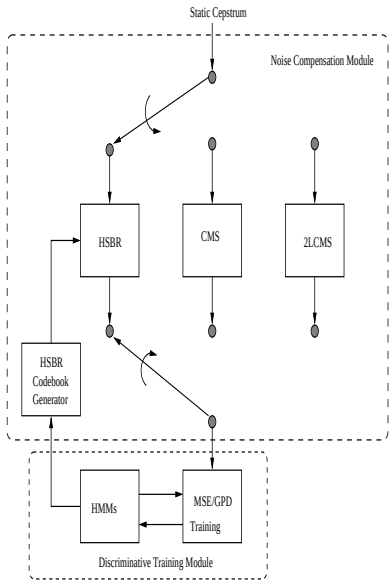

Figure 2: Block diagram of noise compensated MSE training.

- Determine the maximum frame energy E_{max} and minimum frame energy E_{min} for every utterence.
- Separating the frames of current utterance into two classes: if $E_t < \alpha \times E_{max} + (1-\alpha)E_{min}$, then the frame t belongs to class-I (silence class), else to class-II (speech class), where α is a constant determined by a fast experiment. In all the following experiments, we tend to choose the same α value (α is set to 0.2 in our current study), for every database.
- The background and the speech cepstral mean vectors are calculated for the whole uttereance.
- Finally the normalized cepstral features for each frame are computed by subtracting them by their respective cepstral means.

The HSBR method is a blind equalization scheme and it also aims at reducing the acoustic mismatch between the training and various testing conditions [4]. It incorporates the signal bias removal approach with hierarchical clustering method where the size of the codebook is dynamically expanded for signal bias compensation. In other words, the data is hierarchically divided into smaller clusters at each level after computing a separete bias for each cluster in the previous level. Typically, the codebook is generated from the clustering of all acoustic feature vectors in the training set. In our experiments, the codebook used in HSBR is generated from the mean vectors of the HMMs and the maximum codebook size is set to 4. Noise compensation is important not only for testing but also for training. In this work, the above noise compensation procedure is integrated into MSE training as illustrated in Fig. 2.

4. Mandarin Speech Database

The Mandarin connected-digit database contains speech of 125 speakers from broad range of accents and dialect regions within Shanghai, China [12]. The data distribution of the training and

Databases	Training Strings	Testing Strings
DB1	1145	220
DB2	413	76
DB3	406	78
DB4	412	74
DB5	402	70
DB6	517	91
DB7	515	91
DB8	609	116
DB9	414	74
DB10	395	78
DB11	405	72
DB12	398	68
DB16	609	107
Total	6640	1215

Table 1: Distributions of spoken digit strings among the training and testing sets of the connected-digit database.

testing set is shown in Table 1. The Mandarin database contains digits zero through nine and another pronunciation for one (denoted as A). Usually, one would pronounce digit one as one or A, but can't utter both in one utterance. Thus for each digit string, at most 10 different pronunciations will be uttered as illustrated in Fig. 3. All recordings were made over a dialed-up telephone line, using a standard telephone handset and the environment is divided into home, office, street and shop. Speech was digitized at 8 KHz rate using a 16 bit A/D and stored in raw PCM format. The data was then checked for correct word inputs by listening to each recorded string. The training database consists of digit string lengths range from 1 to 16 digits that were spoken by 105 speakers (49 male and 56 female), each speaker provided about 65 digit strings, for a total of 6640 valid strings. The testing database has 20 speakers (10 male and 10 female), each speaker provided about 65 digit strings, and only the valid digit strings were selected for a total of 1215 strings. None of the speakers in the testing database appeared in the training databases.

5. Experimental Results

The recognizer feature set consists of 39 features that includes the 12 cepstral coefficients, log-energies, their first and second order derivatives [9]. The energy feature is batch normalized during training and testing [2]. Each feature vector is passed to the recognizer which models each word in the vocabulary by a set of left-to-right continuous mixture density HMM using context-dependent head-body-tail models. In this study, we model all possible inter-word coarticulation by ignoring the tones, resulting in a total of 275 context-dependent sub-word models. Each model is represented with 3 or 4 states, each having multiples of 4 mixture components [1]. Silence is modeled with a single state model having 32 mixture components. Training included updating all the parameters of the model, namely, means, variances and mixture gains using maximum-likelihood estimation (MLE) followed by six epochs of minimum string error (MSE) training to further refine the estimate of the parameters [4]. The number of competing string models was set to four, the step length was set to one and the length of the input digit strings are assumed to be unknown during the model training and a constrained unknown-length grammar as shown in Fig. 3 is used during testing. Each training utterance is noise compensated prior to being used in training and testing [4]. Penalties based on duration distributions are also applied to the likelihood score.

Before we present the full set of experimental results, we

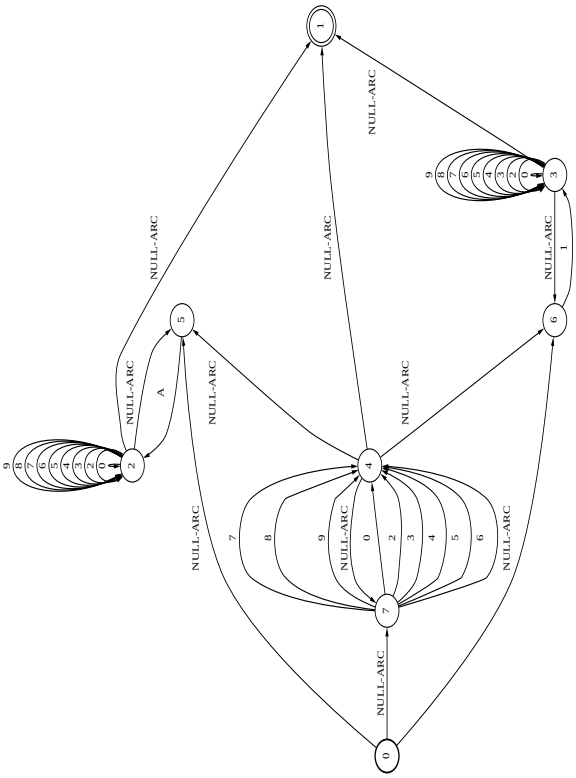

Figure 3: State diagram for constrained-unknown-length (CUL) grammar.

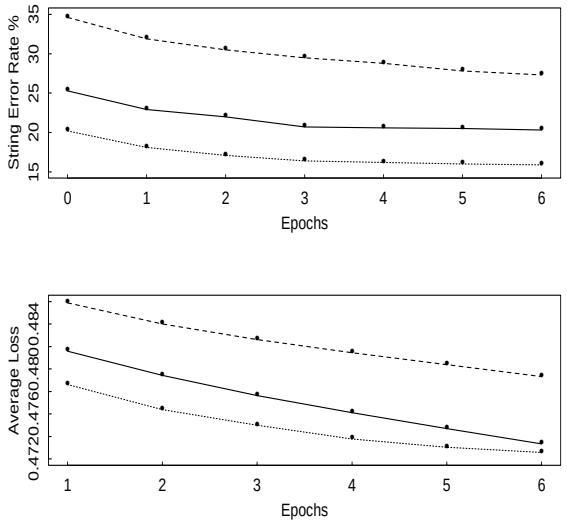

Figure 4: Convergence characteristics of the MSE training procedure: solid-line is MFCC, dashed-line is LPCC and dotted line is AFCC.

first present the convergence characteristics of the MSE training. Fig. 4 shows the connected-digit string error rate of the test set (top plot) and the average loss (bottom plot) of the training set for three different recognizers as a function of the training epoch number. The solid lines are associated with MFCC-based HMM, the dotted lines with HMM generated using AFCC features, and the dashed lines with LPCC-based models. We observed that the string error rate and average string loss both monotonically decrease with the training epoch, with both measures reaching their respective asymptotic values after six epochs of the MSE training. The average loss decreases faster for the AFCC than for the MFCC and LPCC. This indicates that the original objective set out for minimizing the string error via the MSE training is accomplished and that the MSE training is more effective for the AFCC than the LPCC and MFCC features. Several sets of experiments were

Type of Features	MLE		MSE	
	Wd_Er	St_Er	Wd_Er	St_Er
LPCC	7.8%	34.6%	5.9%	27.3%
MFCC	5.4%	25.3%	4.3%	20.3%
AFCC	4.4%	20.2%	3.4%	15.9%

Table 2: Word error rate (Wd_Er) and string error rate (St_Er) for an unknown-length connected-digit recognition task using the conventioanl MLE and MSE training methods as a function of feature type.

Type of Features	MLE		MSE	
	Wd_Er	St_Er	Wd_Er	St_Er
Baseline	5.8%	25.6%	4.4%	19.2%
HSBR	5.1%	23.1%	3.6%	16.9%
CMS	4.4%	20.2%	3.4%	15.9%
2LCMS	4.1%	18.7%	3.2%	14.7%

Table 3: Word error rate (Wd_Er) and string error rate (St_Er) for an unknown-length AFCC based connected-digit recognition task using the conventional MLE and MSE training methods as a function of noise compensation type.

run to evaluate the Mandarin connected-digit recognizer using three types of acoustic features (LPCC, MFCC and AFCC) and three types of noise compensation methods (HSBR, CMS and 2LCMS). Tables 2 and 3 illustrate four important results. First, under all conditions, the MSE training is superior to the MLE training and achieves an average of 25-40% string and word error rate reduction across all types of speech models. Second, the MLE trained MFCC model is better than the MSE trained LPCC model and the MLE trained AFCC model is better than the MSE trained MFCC model. Thirdly, MFCC performs considerably better than LPCC (an improvement of almost 25string error rate reduction). Finally, The MSE trained AFCC yields about 42% (drops from 27.3% to 15.9%) and 22% (decreases from 20.3% to 15.9%) of string error rate reductions in comparison with LPCC and MFCC systems. Clearly, the 2LCMS with AFCC yielded an additional string error rate reduction of 24%, 13%, and 8% when compared to that of the baseline (without any noise compensation), HSBR, and CMS systems.

6. Conclusions

In this study, our focus is on acoustic features and noise compensation for the purpose of improving the recognition performance in adverse ambient conditions. Three different features based on human voice production, perception and auditory systems for the analysis of speech are described and examined for Mandarin speech recognition. The convergence property of the MSE training is shown to be more effective for the AFCC than the LPCC and MFCC features, and test results also supports the superiority of AFCC front-end feature. Further, we explored various noise compensation techniques such as CMS, 2LCMS and HSBR. The greatest word and string accuracies with unknown-length decoding have been obtained when AFCC is used in conjunction with 2LCMS and with MSE trained models (about 96.8% and 85.3%).

Acknowledgements

The author would like to thank Dr. Qi Li of Bell Labs for sharing the auditory based front-end feature extraction software.

7. References

[1] R. Chengalvarayan, "Hybrid HMM architectures for robust speech recognition and language identification," *Proc. Systemics, Cybernetics and Informatics*, Vol. 6, pp. 5-8, 2000.

[2] R. Chengalvarayan, "Hierarchical subband linear predictive cepstral (HSLPC) features for HMM-based speech recognition", *Proc. ICASSP*, pp. 409-412, 1999.

[3] R. Chengalvarayan, "Look-a-head sequential feature vector normalization for noisy speech recognition", *Proc. ICSLP*, pp. Vol. 4, 524-527, 2000.

[4] W. Chou, M. Rahim and E. Buhrke, "Signal conditioned minimum error rate training", *Proc. EUROSPEECH*, pp. 495-498, 1995.

[5] S. Furui, "Recent advances in robust speech recognition", *Proc. ESCA Workshop on robust speech recognition*, pp. 11-20, Pont-a-Mousson, France, 1997.

[6] C. N. Jacobsen and J. G. Wilpon, "Automatic recognition of Danish natural numbers for telephone applications", *Proc. ICASSP*, pp. 459-462, 1996.

[7] H. Jiang, F. Soong and C. H. Lee, "Hierarchical stochastic feature matching for robust speech recognition", *Proc. ICASSP*, 2001.

[8] D. Langmann, A. Fischer, F. Wuppermann, R. Haeb-Umbach and T. Eisele, "Acoustic front-ends for speaker-independent digit recognition in car environments", *Proc. EUROSPEECH*, pp. 2571-2574, 1997.

[9] Y. Lee and L. S. Lee, "Continuous hidden Markov models integrating transitional and instantaneous features for Mandarin syllable recognition", *Computer Speech and Language*, No. 7, pp. 247-263, 1993.

[10] Q. Li, F. K. Soong and O. Siohan, "A high-performance auditory feature for robust speech recognition", *Proc. ICSLP*, Vol. 3, pp. 51-54, 2000.

[11] J. Picone, "Signal modeling techniques in speech recognition", *Proc. IEEE*, Vol. 79, No. 4, pp. 1215-1247, 1993.

[12] F. Zhao, P. Raghavan, S. K. Gupta and Z. Lu, "Automatic speech recognition in Mandarin for embedded platforms", *Proc. ICSLP*, Vol. 2, pp. 815-818, 2000.