Abstract
Due to the mismatch between training and operating conditions, speech recognition systems often exhibit dramatic performance degradation when they are practically used in real-world environments. Various techniques based on noise resistance, speech enhancement and model compensation approaches have been widely used to improve the performance of speech recognition in noise. Parallel Model Combination (PMC) is one of the most popular techniques and has been proved powerful in compensating recognition models, so that they reflect speech in noisy acoustic environments. However, studies have shown that some assumptions and approximations made in the PMC, primarily in the domain transformation and parameter combination processes are not accurate in certain situations, thus restricting the achievement of better performance. This research suggests using Maximum Likelihood Spectral Transformation (MLST) as the adaptation technique to further improve the performance of PMC. MLST is a transformation-based technique. It has some advantages over other adaptation techniques like Maximum Likelihood Linear Regression (MLLR) and Maximum A Posterior (MAP). MLST requires only a small amount of adaptation data and the adaptation process is computationally inexpensive. The proposed method is denoted as Adaptive Parallel Model Combination (APMC).

Keywords
Parallel Model Combination (PMC), Maximum Likelihood Spectral Transformation (MLST), model compensation, adaptation technique, noisy speech recognition

1. Introduction
Nowadays, the performance of those state-of-the-art speech recognition systems has been significantly increased. Speech can be recognized with accuracy higher than 90% using the state-of-the-art speech recognition systems (Hunter, 2002; Furui, 2005). Anyway, this is often the result gained in controlled environments, which are usually quiet environments. In real world application, it is rarely quiet and there is also little control over the acoustic environment. This is especially true in these days since speech recognition has been widely used on handheld devices like PDA and cellular phone. The portability of these devices means the users use speech recognition not only in the quiet office, but also in shopping malls, in cars, on the streets, etc. According to some estimates, there is a 20% - 50% decrease of recognition rates when speech recognition is implemented in a natural field environment (Oviatt, 2000; Rose, 2004).

Figure 1: Main causes of speech recognition performance degradation in real-world application

Figure 1 summarizes the main causes that contribute to the degradation of speech recognition performance in real-world application. The principal cause of dramatic performance degradation is the distinct difference between training and operating condition. This environmental difference is caused by noises, which are the sounds that are transmitted to a recognition system but are not part of the meaningful input signal. The predominant kind of noise is the background noise, which is produced at the speaking environment and being input to the recognition
system along with the speech. Background noise is superimposed upon the speech input, i.e. it is additive. When speech signals are transformed and transmitted through some speech channels, channel noise is also induced on the spoken input. Two primary channels in speech recognition are microphones that are used during recording and telephone channels that serve as the communication channels. Environmental noises are different in their characteristics. Basically, signal-to-noise ratio (SNR) becomes an objective measure to estimate the amount of noise present with a speech signal. Studies show that performance of a recognizer is rather uniform for SNRs greater than 25dB but there is a very steep degradation as the noise level increases. In training environment or laboratory, SNRs may reach 90dB (but 50dB would be more typical). Unfortunately, most real-world environments have poor SNRs. For instance, in factory or subway, the SNRs are usually less than 5dB (Markowitz, 1996). This significant gap causes most recognizers to perform poorly in operating condition. Besides of SNRs which greatly related to the loudness of the noise, the accuracy of speech recognition also depends on the nature of the noise. For instance, non-speech noise is easier to deal with, compared to background conversation or interfering speech that make discrimination difficult. The variability of the noises, or how much the noises change, is another differentiating characteristic of noises. Steady or periodic noises (engine noise) are easier to identify and eliminated compared to impulsive noises (e.g. machine gun) because of their consistency (Junqua and Haton, 1996). In the case of channel distortion, characteristics of noises may also depend on the types of speech channel and the quality of the channel.

Besides of environmental disturbances, the speakers also introduce some difficulties to the task of speech recognition. During recognition, speech noise (e.g. lip smacks) and non-communicative sounds (e.g. speaker’s hesitation, “Uh...”) are produced by the speakers. Another issue is the Lombard effect which describes the changes that occur to speech when speakers attempt to make them heard over noise. Some of the speech and acoustic features affected by a speaker’s attempt to overcome the noise include increased vocal effort, greater duration of words due to increased vowel length, etc. The Lombard effect makes speech easier to hear and understand when the listener is another human, but in speech recognition, it has a negative impact that may decrease the accuracy up to 25% (Markowitz, 1996). These noises and changes deserve more efforts especially in spontaneous speech recognition because spontaneous speech is ill formed and usually includes redundant information such as disfluency, fillers, repetitions, repairs and word fragments (Furui, 2005).

The following sections describe recent progress in increasing performance of noisy speech recognition, with focus given to model compensation using Parallel Model Combination (PMC) method. This paper then discusses about using Maximum Likelihood Spectral Transformation (MLST) as the adaptation technique to improve PMC. Finally, the concluding remarks are given in section 5.

2. Techniques for Dealing with Noise

On the other hand, various techniques have been proposed by researchers to improve noisy speech recognition. These techniques fall into three main categories: noise resistance, speech enhancement and model compensation.

![Figure 2: Noise resistance approach](image)

Noise resistance (Figure 2) focuses on the effect of noise to the parameters of a model. Since the parameters of recognizer are sensitive to disturbances such as noise in the environment, by observing the effects of these disturbances, it is hoped that parameters that are noise resistant can be derived. In other words, the goal of this approach is to choose speech parameters so that the distortions due to noise are minimal. Recent researches of noise resistance techniques focus on developing computational models of the auditory system. Other techniques in this category include RASTA (Relative Spectral Processing), CMN (Cepstral Mean Normalization), linear discriminant analysis, etc.
Speech enhancement (Figure 3) is usually used as a preprocessing step before recognition. The aim of speech enhancement is to restore clean speech signal or estimate the clean speech from the noisy speech. This involves the transformation of noisy speech into an approximation of clean data used in the training environment. This approach is originally introduced for speech quality improvement. In this case, performance can be assessed by means of intelligibility, quality measures, and distortion between clean and recovered speech. When it is used in speech recognition, performance is mainly in terms of recognition accuracy. Depending on the application domain, it is important to select the relevant evaluation measure because some techniques may enhance the speech but reduce the recognition accuracy (Junqua and Haton, 1996). Spectral subtraction, statistical modeling, mapping transformation, etc. are techniques based on speech enhancement.

As mentioned previously, the degradation of performance is mainly caused by difference between the training and operating environment. The most straightforward approach to obtain good recognition models is to collect enough data from the operating environment, and use the data for training. Anyway, this is almost implausible and very time-consuming. As a result, the recent approaches for noisy speech recognition trend to utilize limited data from noisy environment to update the clean speech models (Hung et al., 2001). This approach is called model compensation (Figure 4). Unlike other approaches, model compensation allows the presence of noise in the recognition process. Usually, Hidden Markov Models (HMMs) are used as the framework of this approach. Model Compensation entails the transformation of the parameters of a clean speech model to accommodate noisy speech. The transformation on the model parameters reduces the discrepancy between training and operational conditions, thus decreases the mismatch between the training and testing. Noisy parameters such as mean and variance are usually adapted for the compensation purpose (Junqua and Haton, 1996). Since model compensation modifies the acoustic models in the pattern matching stage instead of the parameterization stage, there is an advantage that no decisions or hypotheses about the speech are necessary and the observed data is unaltered (Gales, 1995). Techniques for model compensation include HMM decomposition, Parallel Model Compensation (PMC), data contamination, Minimum Error Training, adaptation of HMM parameters, etc.

3. Parallel Model Combination

Among the model compensation techniques, PMC has been one of the most popular used techniques. The aim of PMC is to alter the parameters of a set of HMM Based acoustic models, so that they reflect speech spoken in a new acoustic environment. Figure 5 shows the basic process of original PMC method (Gales, 1995). The first step of the process is to transform the inputs, i.e. the clean speech models and noise model, into the linear-spectral domain or log-spectral domain. Transformation is needed because most state-of-the-art speech recognition systems are based on cepstral parameters, but the effects of noise are best modeled in the linear spectral or log-spectral domain. In these domains, the clean speech model and noise model are combined using some mismatch function, which is the function that describes how noise affects the speech parameters. After the model combination, the corrupted speech model is transformed back into the cepstral domain for normal recognition process.
In the MFCC domain, PMC has been shown to produce greater improvements in recognition rate than that of Varga’s HMM decomposition technique (Varga and Moore, 1990). Comparable results were also achieved for non-linear spectral subtraction, but MFCC model combination was predominant because the technique is applicable to a wider range of noise environments and it is not dependent on accurate end-pointing during recognition (Gales and Young, 1992). PMC was also shown to be robust against additive noise for SNRs as low as 0 dB in the cases of stationary and non-stationary noises. In Vaseghi and Milner (1995), good performance has been reported for impulsive noise such as machine gun noise.

The original PMC method is the log-normal PMC (Gales and Young, 1993), which uses log-normal approximation (Gales and Young, 1992) during model parameters estimation. Unfortunately the computation requirements are still relatively high, so a simplified version of PMC that uses log-add approximation (Gales and Young, 1995) is introduced. Anyway, this approach is limited, as it may not be used to compensate the variance parameters.

Another variant of the PMC method is the data-driven PMC (DPMC) (Gales and Young, 1995). In DPMC, the Gaussian integration to estimate the parameters of corrupted speech models is replaced by the generating of corrupted speech observations. That is, a speech observation and a noise observation for a particular pair of speech and noise states are generated, and then combined according to the appropriate mismatch function. Finally, a maximum likelihood estimate of these noisy samples is computed. DPMC is faster and it is also less computationally expensive compared to PMC with comparable performance. However, study showed that the accuracy of DPMC highly depends on the number of samples used to train the noise-corrupted speech model, thus increases the computation complexity while improving the performance (Hung et al., 2001). It is also believed that some assumptions and approximations made in the PMC, primarily in the domain transformation and parameter combination processes, are not accurate enough in certain practical situations. Three new approaches, including the truncated Gaussian approach and the split mixture approach for domain transformation process and the estimated cross-term approach for parameter combination process, are proposed for improvement (Hung et al., 2001). Experiments showed that the proposed approaches are able to provide significant improvements over the original PMC method, especially when the SNR condition is worse. It is believed that the proposed approaches can be further improved if the various parameters of the three approaches can be set differently for different models or different mixtures of the models.

Sarikaya and Hansen (2000) claimed that although PMC is effective, the intense computational complexity limits its use in real-time use. Therefore, they presented a new framework for fast model combination, by using principal component analysis (PCA) to exploit a priori knowledge of the clean speech models and using efficient rectangular DCT (discrete cosine transform) and inverse DCT matrices for domain transformation. Keeping the same performance as original PMC, the new PCA-PMC method has increased the computational speed by a factor of 1.9. A direct adaptation scheme of the cepstral variance using a linear interpolation has also been introduced by Hwang et al. (2001) to avoid the mapping from linear-spectral domain to the cepstral domain. The interpolation weight is effectively a similarity measure of the combined signal to the clean speech. Two methods to compute the weight are introduced by either using an energy ratio or a distance ratio. This method provides a comparable or superior recognition accuracy to the original PMC and a lot of computation resulted from the parameter mapping can be reduced. Therefore, the improvement of this method is twofold. Since a great deal of computational
complexity of PMC lies behind the domain transformation step, Kim et al. (2003) stated the potential use of cepstrum decomposition technique to the PMC framework, so that the corrupted speech models can be obtained by adding the means and variances of clean speech models and estimated noise models.

Another shortcoming of PMC is that, we need to find the silence periods from utterances to sample the noise signal and estimate the parameters of the noise model. However, it may not be possible to obtain sufficient noise samples to estimate reliable noise parameters in the short periods of silence. Furthermore, the obtained noise samples may fail to represent the true statistics of the noise signal if the noise changes instantaneously. Therefore, Chung (2002) suggested using an expectation-maximization (EM)-based adaptation approach for the mean of noise. This method has shown improved results compared with the conventional PMC. Instead of assuming statistical approximations as in the conventional PMC, the various statistical information necessary for the HMM parameter adaptation can be directly estimated by using Baum-Welch algorithm. Chung (2005) proposed a data-driven adaptation method for the HMM parameter compensation. Generally, the method showed improved performance compared to the conventional PMC. Improvement was especially significant at low SNRs.

PMC has also been used along with Mel Frequency Discrete Wavelet Coefficient (MFDWC) features to take advantage of both noise compensation and speech features local in frequency domain (Tufekci et al., 2001). It was shown that the MFDWCs are superior to the MFCCs for speech recognition in adverse environments when they are used in conjunction with the PMC method.

More recently, Novotny et al. (2004) proposed an original method combining log-add PMC with a noise power spectral density estimation based on minimum statistics. The stability of speech vector selection under the influence of various background noises has made the suggested solution a beneficial method. The main advantages of this solution are no extra voice activity detector and a relatively low computational load.

4. Adaptation Using Maximum Likelihood Spectral Transformation on PMC

A large amount of data is often needed in order to obtain sufficient model accuracy during training of acoustic models. But the collection of data is time-consuming and expensive. As an alternative, adaptation techniques can be applied. A lot of efforts have been made for adaptation of HMM parameters using a small amount of adaptation data. Most of them were originally introduced for speaker adaptation, i.e. to adapt the models to a particular speaker or a group of speaker. However, these techniques can also be used to perform environmental adaptation.

Adaptation of HMM parameters can be done using several methods such as Maximum A Posteriori (MAP) (Gauvain and Lee, 1994) and Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995). The benefits of these techniques can be evaluated from two aspects: first, how effective is the adaptation; second, the amount of adaptation data needed for effective adaptation. MAP (sometimes referred to as Bayesian adaptation) involves the use of priori knowledge about the model parameter distribution. Hence, if the priori knowledge can give information about what the parameters of the model are likely to be, it is possible to use limited adaptation data to obtain a satisfactory MAP estimate. One obvious drawback to MAP is that it requires more adaptation data to be effective when compare to MLLR. However, when large amounts of adaptation training data become available, MAP begins to perform better than MLLR. MLLR is a transformation-based technique and it works by computing a set of transformations to alter the means and variances in the initial model, so that the model is more likely to generate the adaptation data. MLLR usually updated only the means and variances because the main differences of speaker and environment are assumed to be characterized by these parameters.

More recently, an incremental adaptation technique is proposed for rapid adaptation, which adapts the recognizers to a particular user or environment as the systems are being used. This technique is called Maximum Likelihood Spectral Transformation (MLST) (Kim and Yook, 2004). MLST has some advantages over both the MLLR and the MAP. First, MLST has small number of parameters to be estimated, so it requires only a small amount of adaptation data. MLST is also a
transformation-based technique but the transformation is done in the linear-spectral domain so that the adaptation process is computationally inexpensive.

Adaptation technique has been proved useful to further increase the performance of PMC method (Gales, 1995; Gales and Woodland, 1996). Hence, in this research, MLST will be used to adapt the parameters of the corrupted speech HMMs produced by conventional PMC. Figure 6 shows the framework of the suggested method, which is denoted as Adaptive Parallel Model Combination (APMC) in this research.

Adaptation: MLST

5. Conclusion
In this paper, various approaches to improve noisy speech recognition have been discussed. Parallel Model Combination (PMC) has been one of the most successful techniques. However, there are several shortcomings on the conventional PMC, primarily in the domain transformation and parameter combination processes. Besides of the extra computational complexity, some assumptions and approximations made in these processes are not accurate in certain situations, thus restricting better performance of PMC. A lot of methods have been proposed to deal with these problems; e.g. using principal component analysis (PCA), truncated Gaussian approach, estimated cross-term approach, etc. In this paper, a method denoted as Adaptive Parallel Model Combination (APMC), which combines PMC and Maximum Likelihood Spectral Transformation (MLST) has been proposed. MLST will be used to adapt the parameters of the corrupted speech HMMs produced by the conventional PMC to further increase the recognition accuracy of PMC method.

Reference


