Phone recognition in critical bands using sub-band temporal modulations

Feipeng Li, Sri Harish Mallidi, Hynek Hermansky

Center for Language and Speech Processing
Human Language Technology Center of Excellence
Department of Electrical and Computer Engineering
Johns Hopkins University, Baltimore MD 21218, USA
fli12@jhmi.edu, mallidi@jhu.edu, hynek@jhu.edu

Abstract
This study investigates a multistream phone recognition system, which consists of 21 parallel sub-systems, each covers two critical bands, and fused by a multi-layer perceptron (MLP) system. Within each band, speech information is encoded by the frequency-domain linear prediction (FDLP) feature, which characterizes the temporal modulation of subband envelope. Two experiments are conducted to determine the optimal parameters for speech features, the maximum temporal modulation $F_{max}$ and the context window length $T$, followed by an experiment to evaluate the robustness of the fused system in noise. Results show that the phone accuracies of sub-systems reach the maximum point at about 500–600ms; they keep increasing monotonically as the maximum frequency of temporal modulation changes from 4 to 40 Hz, where it saturates. Tests of the fused system in babble and subway noise at 15 dB SNR indicate that the multi-stream system is more robust to noise than the single-stream baseline system.

Index Terms: multistream, temporal modulations, phone recognition

1. Introduction

Human beings are much more robust to noise than the machine systems in recognizing speech. A critical characteristic of the human auditory system is that it takes a parallel processing scheme for speech perception. The cochlea consists of about 40 critical bands from 0.3 to 8 kHz. Each critical band is working as an independent channel for speech reception. Corruption of any one channel has little impact on the performance of overall system since noise masking occurs only within a critical band. In the 1920s, Fletcher and his colleagues at Bell Labs investigated the contribution of different frequency bands to human speech perception. It was discovered that the average phone error rate of full-band stimuli is about the same as the product of error rates from 20 articulation bands [1], consistent with the assumption that the critical bands are independent for speech recognition. Fletcher then identified one articulation band to be about two critical bands, i.e., 1 mm along the basilar membrane (BM). Human speech perception can be broadly categorized into three basic steps: First, speech is decomposed into multiple critical bands at the cochlea, which defines the signal-to-noise ratio in each channel; then the speech features are extracted by hundreds of inner hair cells (heavily overlapped auditory filters) within each critical band; next the speech features are assembled in the central auditory system and used for phoneme classification.

Unlike human speech perception which takes a scheme of parallel processing, typical automatic speech recognition (ASR) systems use fullband spectral template to match speech segments. Degradation at one frequency usually affects the entire template and makes the system fragile in noise. To compensate for this degradation, typical ASR systems place a heavy emphasis on word and language models as a method to increase the recognition score, which makes the problem even more complex.

Inspired by the research on human speech perception [1], a multistream speech recognition system is proposed, in which the full frequency is divided into multiple bands to improve noise robustness [9]. The system yields around 50% reduction in word error rate on isolated digits in frequency-selective additive noise [9]. Recently, Athineos et al [10] developed an analysis technique for speech signal, named frequency domain linear prediction (FDLP), which is quite different from conventional short-term spectral analysis. This technique estimates the amplitude modulation of subband signals, similar to human auditory processing.

In this study, we develop a multistream phone recognizer based on the framework of [9]. The full frequency is divided into 21 bands, each covers about 2 critical bands. Temporal envelopes are estimated for individual bands and used for phoneme classification in each stream. The 21 streams are then combined using a fusion algorithm.
The rest of the paper is organized as follows. Sec. 2 describes the multi-stream architecture and the FDLP technique for feature extraction. The phoneme recognition setup is explained in Sec. 3. Sec. 4 compares the performance of the multi-stream system with a single-stream front-end. In Sec. 5, we conclude the study with a discussion of the proposed features.

2. Multistream phone recognizer

In this section, we describe the architecture of the multi-stream phone recognizer and front-end processing to extract robust temporal modulation features.

2.1. System architecture

The proposed multistream system is parallel in nature. A schematic diagram of the system architecture is shown in Figure 1. It consists of 21 bands equally distributed from 0 to 8 kHz on Bark scale. A separate MLP-based phoneme classifier is built for each band. The neural networks are trained on the temporal modulation features extracted from individual bands to estimate the phoneme posterior probabilities. Since each stream only provides marginal information, those outputs need to be combined to generate a more reliable estimation. A good fusion algorithm should be sophisticated enough to take into consideration multiple factors like: robustness of each stream, relevance of a stream to a particular phoneme class etc. Several different fusion approaches, such as, the inverse-entropy approach, Dempster-Shafer’s method, and KarhunenLove transform- Multi Layer Perceptron (KLT-MLP), have been proposed for the combination of phoneme posterior probabilities. In this work we use the KLT-MLP based fusion approach. The 40-dimension posterior probabilities are converted into features by computing logarithm and then decorrelated using KLT, which reduces the dimensionality of the features to 25. Features corresponding to each critical band are stacked. Accordingly the input feature vector to the fusion system has a dimension of $21 \times 25$. A second-stage MLP is trained to estimate the final phoneme posterior probabilities.

2.2. Sub-stream phone recognizer

The sub-stream phoneme recognition system is based on the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [2]. The MLP estimates the posterior probability of phonemes given the acoustic evidence $P(q_t = i|x_t)$, where $q_t$ denotes the phoneme index at frame $t$, $x_t$ denotes the feature vector taken with a window of certain frames. The relation between the posterior probability $P(q_t = i|x_t)$ and the likelihood is given by the Bayes rule. The state transition matrix is fixed with equal probabilities for self and next state transitions. Viterbi algorithm is applied to decode the phoneme sequence.

2.3. Frequency domain linear prediction based modulation features

Linear prediction (LP) analysis of a signal attempts to predict the current sample as a linear combination of past samples. Through the extraction of linear dependence, the original signal is described as a result of passing a temporally uncorrelated (white) excitation sequence through a fixed all-pole digital filter. When LP analysis is applied in time domain, the filter comprises a parametric approximation of its power spectrum. The duality of time and frequency domain means LP can be applied to discrete spectral representation of a signal. This process is called frequency domain linear prediction (FDLP). In a manner similar to parametric representation of power spectrum by time domain linear prediction, FDLP provides a parametric representation of Hilbert envelope of the signal. Fig. 2 illustrates the ability of FDLP to model Hilbert envelope.

Figure 2: Illustration of the all-pole modelling with FDLP. (a) a portion of speech signal, (b) its Hilbert envelope and (c) all-pole model obtained using FDLP.

Long-segments (3 - 5 seconds) of speech are decomposed into critical bands, by windowing the discrete cosine transform (DCT) coefficients. Sub-band temporal

Figure 1: Schematic diagram of a multistream phone recognition system
envelopes are approximated by an all-pole model using FDLP. Stacking these temporal envelopes creates a two-dimensional time-frequency representation of the input signal. The sub-band envelopes are converted into modulation spectral components by computing DCT on segments of envelope with a duration of $T$ ms and a time shift of 10 ms. In each segment, the modulation frequency components greater than $F_m$ are discarded. The corresponding number of DCT coefficients used for features can be calculated by $\lfloor (T + 25) \cdot F_m / 500 + 0.5 \rfloor$, where 25 ms is the duration of individual frames. It is observed that the two parameters $T$ and $F_m$ have significant effect on the performance of the substream phone recognizers.

3. Experiments

Two experiments are conducted to determine the maximum modulation frequency $F_m$ and optimal context window length $T$ for sub-stream systems; then the two identified parameters are applied in the multistream system in the third experiment for the evaluation of noise-robustness. The details of the three experiments are explained in the following subsections.

3.1. Experiment I: context window length $T$

This experiment aims to identify the optimal context window length $T$ for all sub-stream phone recognizers. For each band the context window length $T$ is varied from 100 to 800 ms with the maximum modulation frequency $F_m$ being fixed at 40 Hz.

3.2. Experiment II: maximum modulation frequency $F_m$

This experiment aims to identify the maximum frequency of temporal modulation $F_m$ that contributes to phoneme identification. The context window length $T$ is fixed at 600 ms, and the maximum modulation frequency $F_m$ is varied from 4 to 48 Hz with a step size of 4 Hz.

3.3. Experiment III: noise robustness of the multi-stream system

This experiment aims to assess the noise-robustness of the multi-stream system in noise. The multi-stream system is trained in clean condition and tested in clean and subway noise at 15 dB SNR. The performance is compared with that of a single-stream system using PLP feature as the front-end.

All experiments are performed on TIMIT database containing speech sampled at 16000 Hz. The training data consists of 3000 utterances from 375 speakers, cross-validation data set consists of 696 utterances from 87 speakers and the test data set consists of 1344 utterances from 168 speakers. The ‘sa’ dialect sentences are removed from the experiments. The TIMIT database, which is hand-labeled using 61 labels is mapped to the standard set of 39 phonemes.

4. Results and Discussion

In this section, we first present the results of two experiments on maximum modulation frequency $F_m$ and optimal context window length $T$; then we compare the performance of multistream system with single-stream baseline systems under noisy conditions.

![Figure 3: Effect of maximum modulation frequency $F_m$ on phoneme recognition, with context window length $T$ being fixed at 600 ms, in clean condition](image)

4.1. Maximum modulation frequency $F_m$

Figure 3 depicts the phone accuracy of the 7, 13, and 17th band as a function of maximum modulation frequency $F_m$ in clean condition. The performance of all three subband phone recognition systems climb dramatically as $F_m$ increases from 4 to 12 Hz; then the slope of the curves drop immediately, suggesting that amplitude modulation below 12 Hz is critical for speech recognition. From $F_m=12$ to 24 Hz the phone accuracies of the 7th, 13th, and 17th bands increase by 3.2, 3.8, and 3.3% absolute respectively, suggesting that amplitude modulation within 12 to 24 Hz is significant for phoneme classification. After that the phone accuracies of subband phone recognition systems keep increasing slowly as $F_m$ changes from 24 to 44 Hz, where the subband phone recognition systems saturate in performance. The 7, 13, and 17th band, selected from the low, middle, and high frequency range respectively, all generate the same results.

4.2. Context window length $T$

Figure 4 depicts the phone accuracy of the 7, 13, and 17th band as a function of context window length $T$ in clean
condition. When the context window is shorter than 200 ms, all subband phone recognition systems are significantly affected with the phone accuracies of the 7, 13, and 17th band drop by 2.2, 1.4, and 0.94% absolute respectively, suggesting that 200 ms, which is about the average length of syllables, is the critical context for phoneme classification. Beyond the critical context, most subband systems show steady yet slow increase, as the context $T$ changes from 200 to 800 ms. All subband systems reach maximum when the duration of context is somewhere around $T = 500$ ms.

![Figure 4: Effect of context window length $T$ on phoneme recognition, with maximum modulation frequency $F_m$ being fixed at 40 Hz, in clean condition](image)

**4.3. Noise robustness of fused system**

Figure 5 compares the phone accuracy of the fused multistream system with a single-stream system that use perceptual linear prediction (PLP) [6] feature as the front-end. The two systems are comparable in clean conditions. When the speech is corrupted with babble and subway noises at 15 dB SNR, the phone accuracy of the multi-stream is significantly better than that of the single-stream system.

In addition, the multi-stream system allows to choose higher temporal context and modulation frequency compared to single-stream system. This is because, multi-stream system operates on features extracted from individual sub-bands which is considerably smaller in dimension as compared to the feature for single-stream system.

**5. Conclusions**

In this study we developed a multistream phone recognition system that consists of 21 sub-systems, each covers two critical bands, and fused by a multi-layer perceptron (MLP) system. Each of these 21 sub-systems is trained on temporal modulation features extracted by using FDLP to individual sub-bands. Two experiments are conducted to determine the context window length ($T$) and maximum modulation frequency ($F_m$) that optimize the performance of individual sub-systems. Results show that the phone accuracies of sub-systems reach the maximum at around $T = 600$ ms and $F_m = 40$ Hz. These choices are used to build a multi-stream system and tested on babble and subway noises at 15 dB SNR. Compared to the single-stream baseline system, the proposed multi-stream system is more robust to noise.

**6. References**