Phoneme recognition in critical bands based on subband temporal modulations

The Center For Language and Speech Processing
at the Johns Hopkins University

Feipeng Li, Sri Harish Mallidi, Hynek Hermansky
CLSP, Johns Hopkins University, Baltimore, MD 21218, USA
fli12@jhmi.edu, mallidi@jhu.edu, hynek@jhu.edu

1 Introduction

Human speech perception is robust to noise because it takes a parallel processing scheme.

- Cochlea modeled as an array of auditory filter; Acoustic signal gets masked by noise only when the two falling within the same filter simultaneously
- Fletcher and his colleagues discovered that fullband phoneme recognition error is equal to the product of error rates from 20 articulation bands (one articulation band ≈ two critical bands) $e = e_1 e_2 \dots e_{20}$, i.e., corruption in one band has little effect on the overall system.

Therefore, we propose a multistream phoneme recognizer (Fig. 1).

A. Multistream ASR

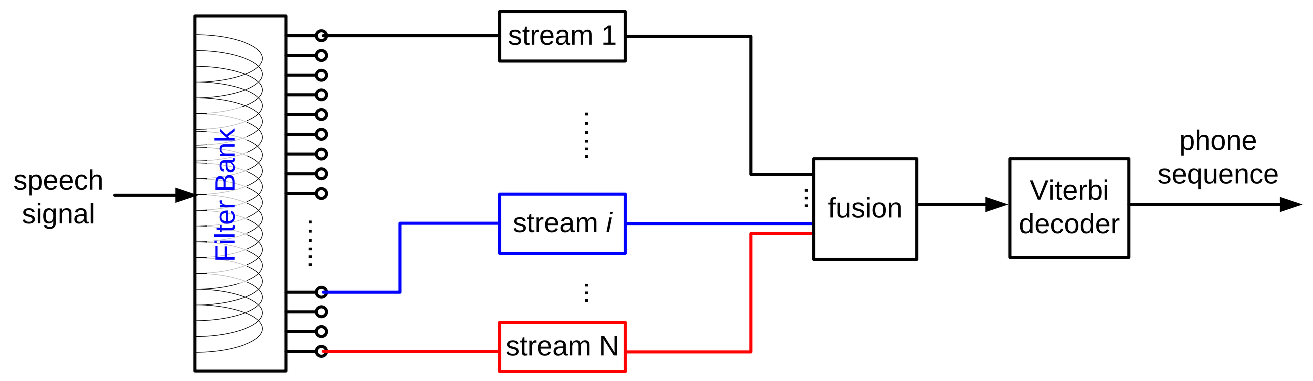

Figure 1: Block diagram of a multistream phoneme recognition system

- Full frequency range divided into 21 stream (2 critical bands/stream); Each stream has an independent MLP-based phoneme classifier, trained on the subband frequency-domain linear prediction modulation (FDLPm) feature.
- Logarithms of posterior probabilities of 21 streams are decorrelated by using Karhunen-loeve transform (KLT), concatenated, and used as feature for a second-stage MLP for fusion.

B. FDLPm feature

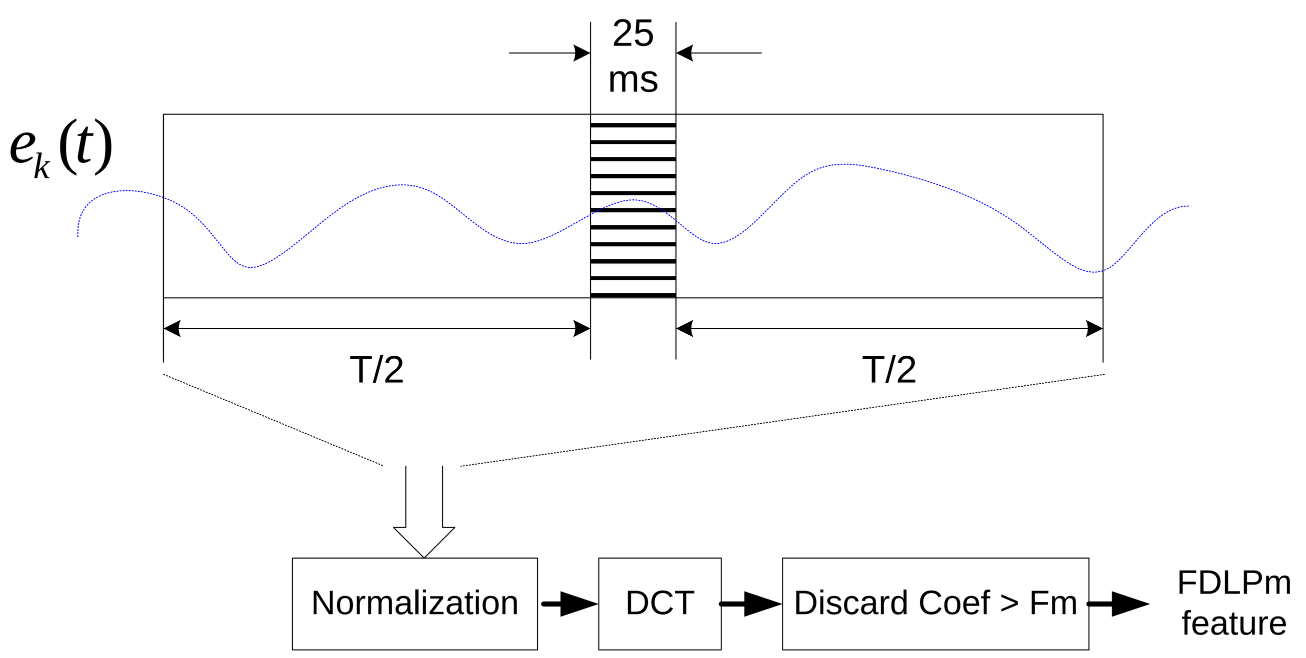

Figure 2: Illustration of FDLPm feature for the subband envelope of the k^{th} stream

Questions:

- What is the optimum bandwidth BW for the auditory filterbank?
- What is the optimum context window T for FDLPm feature?
- What is the optimum temporal modulation frequency F_m for the selection of DCT coefficients?

2 Experiments

Exp. 1: To determine optimum filter bandwidth BW

- Frequency ranges (560,1278)Hz
- Bandwidth BW increases from 0.25 to 3.75ERB with a step size of 0.25 ERB
- Number of filters/Bark = [1,2,3,4]
- Context window T fixed at 200ms and F_m fixed at 35 Hz

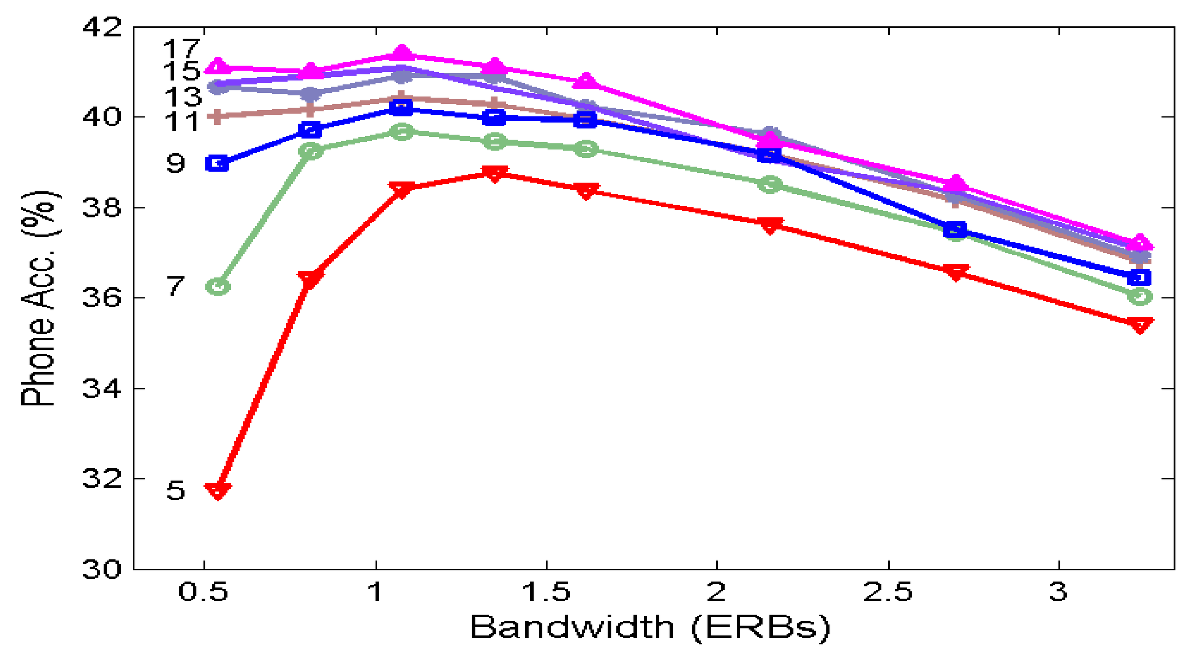

Figure 3: BW = 1 ERB generally produces the best performance

Exp. 2: To determine optimum context window length T

- Context window T increases from 100ms to 800 ms
- Maximum modulation frequency F_m fixed at 40 Hz

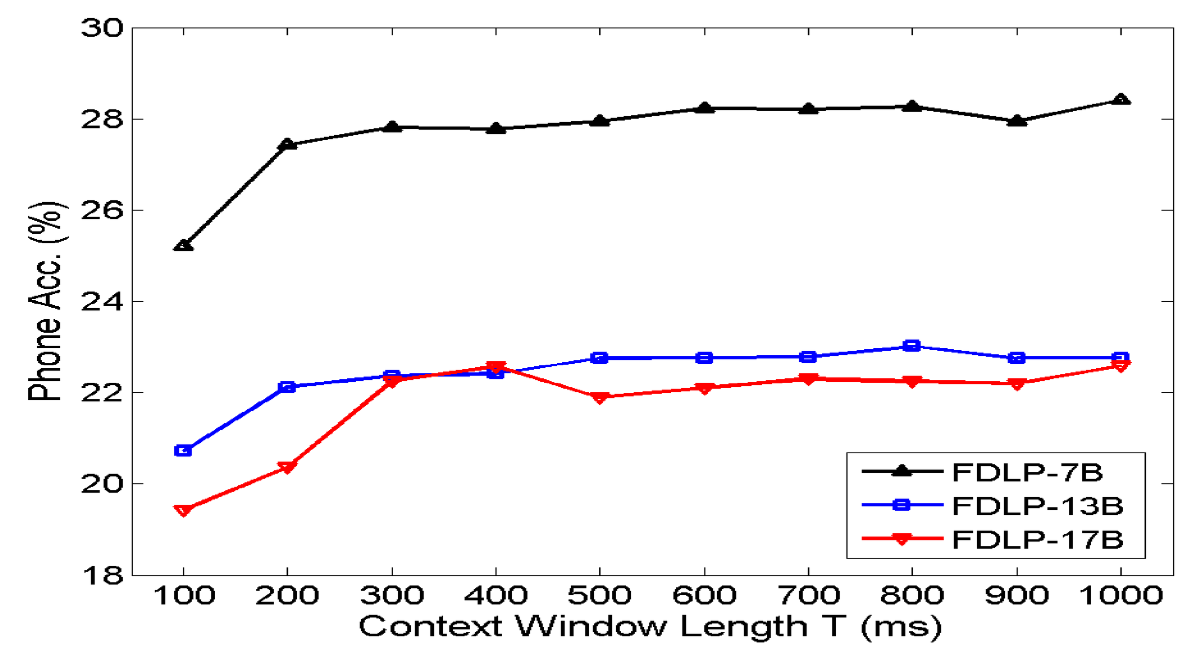

Figure 4: Most subband phoneme recognition systems are significantly affected when the context window T is shorter than 200 ms; Optimum duration of context $T =$ is around 300 ms

Exp. 3: To determine optimum maximum modulation frequency F_m

- Maximum modulation frequency F_m increases from 4 to 48Hz with a step size of 4 Hz
- Context window length T fixed at 600 ms

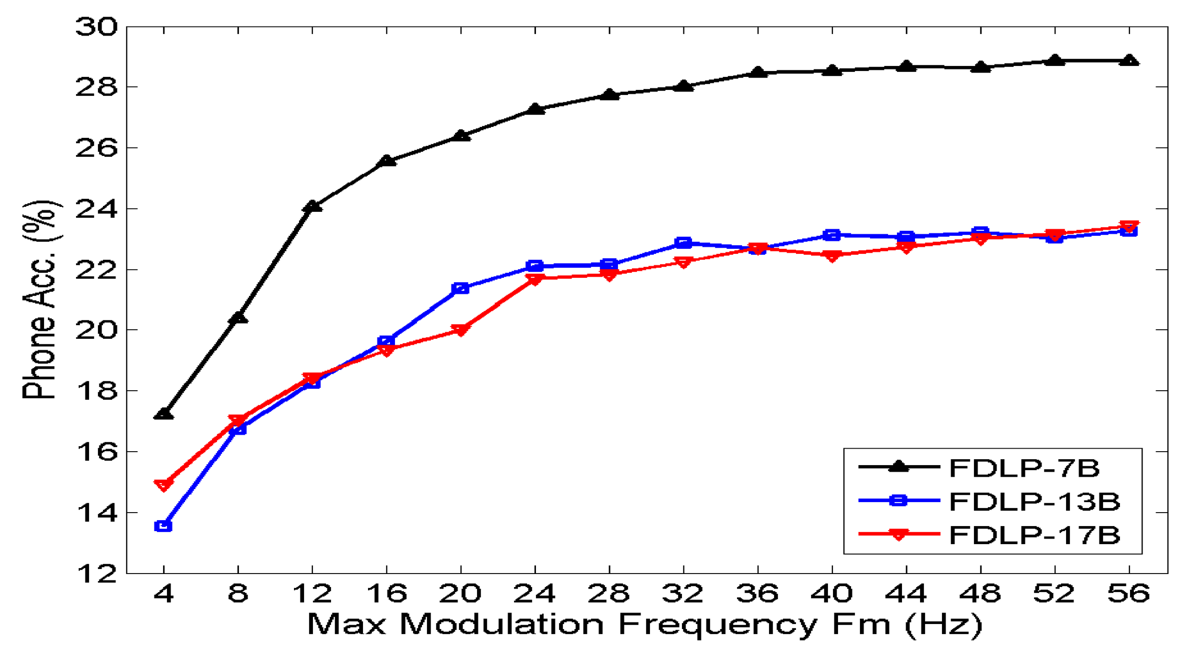

Figure 5: phone acc. of subband systems climb dramatically as F_m increases from 4 to 12 Hz, suggesting that amplitude modulation of 12 Hz is critical for phoneme recognition; optimum F_m is around 32 Hz

3 Noise robustness of multistream ASR

- Multi-stream ASR (optimum T and F_m and $BW \approx 2.5$ ERB) trained in clean conditions and test in clean and subway noise at 15 dB SNR

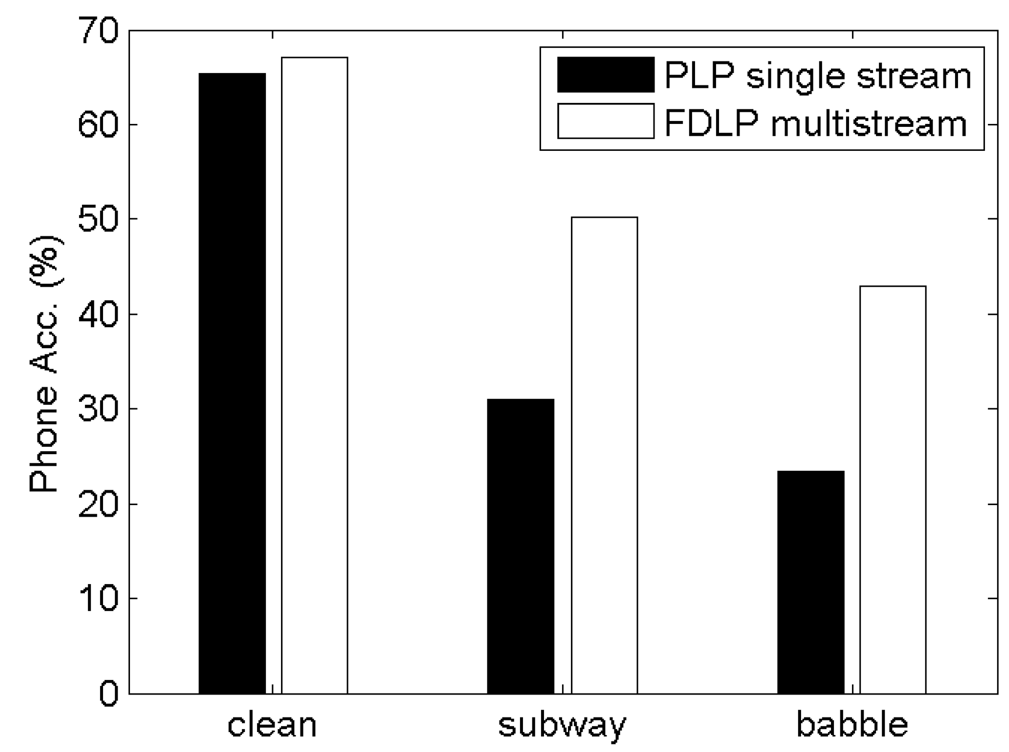

Figure 6: The two systems are comparable in clean condition. The multistream ASR outperform single-stream ASR significantly when the speech is corrupted with babble and subway noises at 15 dB SNR

4 Summary

- We proposed 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.
- Multistream ASR reaches the maximum performance when $T = 300$ms, $F_m = 32$Hz, and $BW = 1$ERB respectively.
- Multistream ASR out-performs the single-steam baseline system in babble and subway noise at 15 dB SNR.