SUBBAND HYBRID FEATURE FOR MULTI-STREAM SPEECH RECOGNITION

Feipeng Li

Center for Language and Speech Processing
Johns Hopkins University, Baltimore, MD, 21218

ABSTRACT

A subband hybrid (SBH) feature is developed for multi-stream (MS) speech recognition. The fullband speech signal is decomposed into multiple subbands, each covers about 3 Bark along the frequency. Speech signal is analyzed by a high-resolution filterbank of 4 filters/Bark and a low-resolution filterbank of 2 filters/Bark to facilitate the representation of both short-term spectral modulation and long-term temporal modulation within a frequency subband. Experiments on TIMIT corpus for English and RATS corpus for Arabic Levantine show that the SBH feature significantly enhances the amount of information being extracted from individual subbands. The MS system with performance monitor achieves a substantial gain in performance over the single-stream baseline.

Index Terms: noise robustness, multi-stream speech recognition, subband feature

1. INTRODUCTION

The human auditory system takes a parallel scheme for speech perception that is highly reliable under realistic environments [5, 1, 10]. To emulate the parallel processing in human speech perception, a prototype multi-stream (MS) phoneme recognition system is proposed in [15], and then upgraded in several major aspects in [17]. The basic idea is to decompose the fullband speech into multiple subbands, each acting as an independent channel for speech recognition. Since stationary white noise is rare in realistic environments, it is assumed that at any time at least one or two of those channels are still functioning properly. For simplicity, no context (i.e., lexical, syntactic, and semantic) information is utilized.

The MS system is superior to conventional single-stream (SS) system in two aspects. First, noise is contained in narrow frequency bands, hence corruption of one channel has little effect on the overall system. Second, the total error rate of speech recognition can be effectively reduced by representing the subband speech with a more informative feature.

Past studies on multi-stream speech recognition [15, 14, 17] were focused on how to fuse the information from multiple subbands. There were few studies on the acoustic features for parallel processing. Typical short-term spectral feature, such as PLP and MFCC, is inappropriate for MS speech recognition because the spectral-to-cepstral transform violates the band-independence constraint for parallel processing. The long-term temporal modulation FDLPm feature [6], which characterizes the peaks of subband envelopes, is parallel by nature, but it is optimized for single-stream systems.

In our previous MS system [17] the speech features were derived simply by splitting the FDLPm feature into multiple segments. Our recent study show that a multi-resolution filterbank of 2.5 and 1 Bark bandwidth significantly out-performs the filterbank of 2.5 Bark uniform bandwidth in FDLPm. In addition, a spectral sampling rate of 4 filters per Bark is generally better than 1 filter per Bark for subband phoneme recognition [11].

In this study we develop a subband hybrid (SBH) feature for multi-stream speech recognition with an aim to boost the performance in individual subbands. The SBH feature represents both long-term temporal modulation and short-term spectral modulation within a frequency subband.

2. MULTI-STREAM SPEECH RECOGNITION

Figure 1 depicts the block diagram of the MS phoneme recognition system. It takes the Hidden Markov Model - Artificial Neural Network (HMM-ANN) paradigm [4]. The full-band speech is decomposed into multiple (7 or 5 for wide/narrow-band speech) independent bands, named *band-limited streams*, and then encoded by the SBH feature for phoneme classification. Next, the information from multiple bands are integrated by a set of neural nets to form many combinations, named *processing streams*. A performance monitor is used to evaluate the quality of posteriors for each processing stream. The top N best posteriors are averaged to produce a more reliable estimation. Viterbi algorithm is applied to decode the phone sequence. It is assumed that all phonemes

This work was supported by the Defense Advanced Research Projects Agency (DARPA) RATS project D10PC0015 and Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, IARPA, DoD/ARL, or the U.S. Government.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

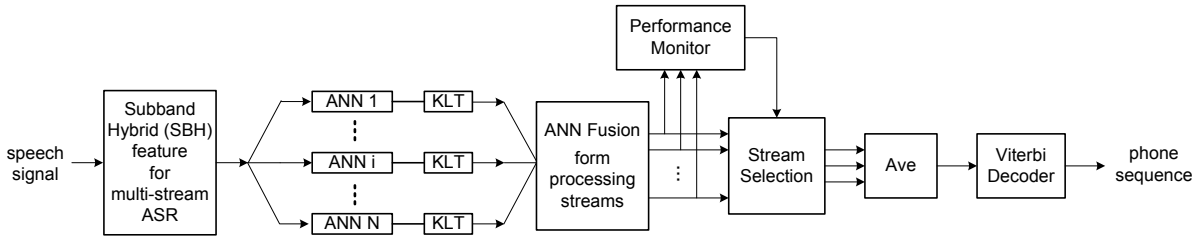


Fig. 1. Block diagram of multi-stream phoneme recognition system

have equal prior probabilities. The state transition matrix is fixed with equal probabilities for self and next state transitions.

Within each subband a multi-layer perceptron (MLP) is trained to estimate the posterior probability of a phoneme given the partial acoustic evidence. The size of the neural net is dependent on the number of training patterns. The dimensionality of the linear output of MLP is reduced to 25 by applying the Karhunen-Loeve Transform (KLT).

The posterior features from band-limited streams are integrated by a set of MLPs trained to fuse each of all possible combinations of band-limited streams (31/127 combinations for 5/7 subbands). A performance monitor named the M measure [7] is used to evaluate the reliability of individual processing streams based on the posterior probability produced by the fusion MLPs. For a given window of T consecutive frames the M measure is defined as

$$M(\Delta t) = \frac{\sum_{t=0}^{T-\Delta t} D(P_t, P_{t+\Delta t})}{T - \Delta t} \qquad (1)$$

where D is the symmetric KL divergence between two distributions of posterior probability, P_t and $P_{t+\Delta t}$, separated by Δt along time. For each utterance, the M measure is computed for all 31/127 processing streams. Streams that are more reliable usually have higher M values. The posterior probabilities of top N best processing streams are averaged to produce a more reliable estimation.

3. SUBBAND HYBRID FEATURE

The schematic diagram of the SBH feature for multi-stream speech recognition is depicted in Fig. 2.

3.1. Speech Analysis

The time-domain speech signal is transformed into frequency domain by using a discrete cosine transform (DCT) transform ($\geq$ 4000 frequency points). Frequency-domain linear prediction (FDLP) [2, 6] is used to produce two spectro-temporal representations for the short-term spectral feature and long-term temporal feature within each frequency subband. Two filterbanks, a high-resolution (hi-res) W_H and a low-resolution (lo-res) W_L, are employed to divide the DCT coefficients into multiple bands. Fig. 3 show the hi/lo-res filterbanks $W^{H/L}$. W^H has a bandwidth of $B^H = 1.0$ Bark with a spectral sampling rate of 4 filters per Bark. W^L has a bandwidth of $B^L = 2.5$ Bark with a spectral sampling rate of 2 filters per Bark. For a band-limited stream of about 3 Bark, the hi/lo-res filterbanks have 12 and 6 filters respectively. Speech signal is analyzed by multiplying the DCT coefficients with a set of cosine windows (Eq. 2), as depicted in Fig. 3.

$$W_m^{H/L}(f) = 0.5 + 0.5\cos(\pi(f - f_m)/B^{H/L}) \qquad (2)$$

where f_m and B are the center frequency and 6dB filter bandwidth (i.e., amplitude = 0.5) of the m_{th} window (filter), both on Bark scale. Since the cosine windows are defined on Bark scale, this process also converts the linear frequency into auditory frequency. Next, the DCT coefficients are divided into multiple subbands (5 and 7 for narrow/wide speech respectively). Two FDLP spectrums are generated based on the outputs of hi/lo-res filterbanks respectively. The hi-res spectrum is used for short-term spectral feature, while the lo-res spectrum is interleaved with the odd bands of hi-res spectrum and used for the computation of long-term temporal feature.

3.2. Short-term Spectral Modulation Feature

The short-term spectral modulation feature is computed on the basis of hi-res FDLP spectrum, which includes 12 envelopes of band-passed signals. To estimate the short-term spectral amplitude, the FDLP spectrum is divided into short frames of 25 ms with a step size of 10 ms, and then multiplied with a hamming window, followed by averaging along the time axis. Next, the cepstral coefficients of the log spectral amplitude are liftered with a cutoff quefrency of 0.6. The first 9 cepstral coefficients (DC component not included) and their first and second derivatives are used as the short-term spectral modulation feature.

3.3. Long-term Temporal Modulation Feature

The long-term temporal modulation feature includes a static module, which involves a logarithmic operation, and an adaptive module [6], which simulates the adaptive compression of auditory periphery [16]. The outputs from both modules are

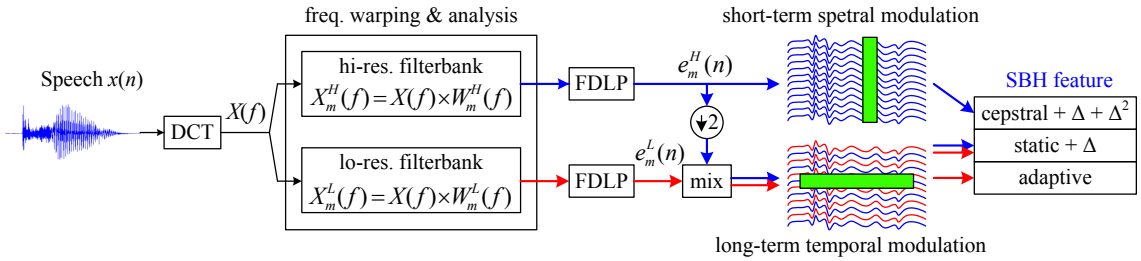


Fig. 2. Schematic diagram of SBH feature for multi-stream speech recognition

divided into short segments of 225 ms (i.e., the current frame plus 10 frames before and after) with a step size of 10 ms. Then the framed signals are transformed into the modulation frequency domain by a DCT transform. The first 18 coefficients (corresponding to a cutoff modulation frequency of 40 Hz) of all 12 bands, together with frequency differentiation of lo-res spectrum are concatenated to form the long-term temporal modulation feature.

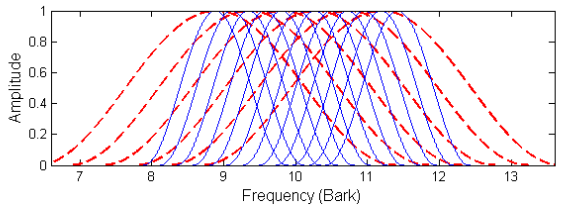


Fig. 3. Hi-res filterbank W^H (solid curve) and lo-res filterbank W^L (dashed curve) for subband feature

4. EXPERIMENTS

The SBH feature is tested for multi-stream phoneme recognition on both TIMIT corpus for American English and RATS corpus for Arabic Levantine.

4.1. Phoneme Recognition Systems

The MS system is described in Sec. 2. For comparison, we also tested the single-stream (SS) baseline system using the MFCC, PNCC[9], and FDLPm feature [6]. It has two modules, a MLP trained to discriminate the phonemes based on the acoustic feature, followed by a Viterbi decoder. Both the MS and SS system are built as single-state monophone systems without using any context information.

The MLP has a size of $588\times1000\times40$ for TIMIT and $420\times1000\times1000\times250\times38$ for RATS. In the experiment on TIMIT all neural nets are trained under clean condition. In the other experiment on RATS, neural nets are trained and tested under the same condition (channel) because the 9 channels are distinctive from each other. Neural nets trained on one channel do not generalize across other channels.

4.2. Speech Corpus

TIMIT corpus for American English consists of 5040 sentences, of which 3696 are used for training, and the rest 1344 are used for testing. The target phoneme set contains 40 phonemes. The speech sounds are sampled at 16kHz.
RATS corpus for Arabic Levantine consists of 9 channels, specifically, src for clean speech, and A, B, C, D, E, F, G, H for various types of distortion. Each channel contains about 13–23 hours of speech. The target phoneme set consists of 38 phonemes. The speech sounds are sampled at 8kHz.

4.3. Data Analysis

In the first experiment on TIMIT speech corpus, the comtribution of SBH feature for multi-stream speech recognition is evaluated by comparing the results of multi-stream system using FDLPm feature. Two types of multi-stream systems are included in this study. The **MS** denotes a simple version of multi-stream system that has only one processing stream that covers all subbands. The **MS-PM** denotes the full version of multi-stream system with a performance monitor for stream selection. The **Oracle** performance refers to the results of **MS-PM** system, in which the best stream is manually picked for each utterance. For completeness, the results of multi-stream system are also compared with those of single-stream baseline systems (denoted as **SS**) using MFCC, PNCC, and FDLPm feature. In the second experiment, the results of SBH feature for multi-stream systems are compared with those of single-stream PLP baseline as well as oracle performance.

5. RESULTS

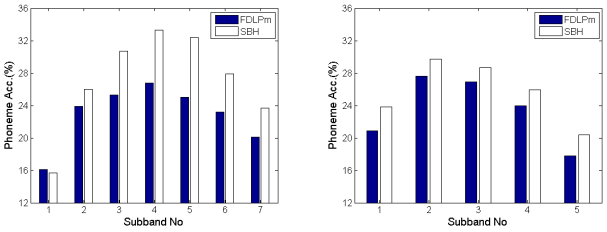
Results show that the SBH feature substantially out-performs the FDLPm feature for almost all frequency bands. Fig. 4 depicts the subband phoneme accuracy, averaged over all conditions (channels), for both TIMIT and RATS. The subband phone sequence is obtained by applying the Viterbi decoder to the posteriors from each subband neural net. For TIMIT database the average phoneme accuracies of SBH feature are about 10–30% relative higher than those of FDLPm feature for subband 2–7, except for subband 1, which is relatively

Table 1. Percent Error Rate (PER) (%) of multi-stream system on TIMIT

Noise	SS			MS		MS-PM		Oracle	
(dB SNR)	MFCC	PNCC	FDLPm	FDLPm	SBH	FDLPm	SBH	FDLPm	SBH
clean	33.50	33.51	31.35	31.27	30.04	29.98	29.45	23.78	24.01
babble (15)	65.14	58.93	57.10	52.80	47.05	49.68	45.16	42.85	39.22
subway (15)	57.42	49.48	46.62	45.15	38.28	40.79	36.28	34.11	30.65
factory1 (10)	74.65	72.05	68.10	69.87	63.74	67.10	61.56	59.91	54.72
restaurant (10)	72.48	67.14	63.14	65.03	58.19	61.61	55.69	55.18	49.38
street (5)	80.98	70.26	67.26	68.47	62.56	65.27	59.37	58.08	52.60
exhall (5)	79.90	74.97	70.67	71.16	69.47	68.67	64.46	61.85	58.57
f16 (0)	77.32	84.31	86.10	85.30	86.54	85.71	83.83	76.78	75.35
car (0)	90.24	52.58	54.32	48.76	40.79	40.24	35.48	34.30	30.14

∗ **SS** – single-stream baseline; **MS** – multi-stream system with only one stream including all subbands; **MS-PM** – multi-stream system with a performance monitor; **Oracle** – multi-stream system with best stream manually picked for every utterance after decoding.

unimportant for speech recognition. For RATS corpus of Arabic Levantine the SBH feature is about 7–15% relative better than the FDLPm feature for all 5 subbands.



(a) TIMIT (b) RATS Arabic

Fig. 4. Subband phoneme accuracy of FDLPm and SBH (averaged over clean and noisy conditions or channels)

The MS system with SBH feature is significantly better than the same system using FDLPm feature for both clean and all noisy conditions (refer to Tab. 1). With performance monitor the percent error rate of MS-PM system is about 10–35% lower than that of the single-stream FDLPm baseline for almost all noisy conditions. The most gain in performance (35%) is achieved for 0dB car noise, which affects only the low frequency subbands. In contrast, the single-stream MFCC baseline totally fails, while the and PNCC baseline suggesting that the MS-PM system is superior for narrow-band noise.

The MS-PM system show marginal gain (2.5%) for f16 noise which has a nearly white spectra and contaminates all subbands. For almost all conditions except for exhall and f16 noise, the use of SBH features brings higher gain in performance than stream selection by using performance monitor, suggesting that the subband feature is critical for the success of multi-stream system.

In the other experiment on RATS Arabic Levantine, the MS-PM system show slight improvement on channel A, B, C, D and about 10–50% increase in phoneme accuracy for channel E, F, G, and H (refer to Tab. 2). Again, the use of SBH features contributes most of the gain in performance. It is hypothesized that the gain of stream selection diminishes for channel A, B, C, D because speech in those channels are corrupted by heavy noise in all 5 subbands.

Table 2. Phoneme Acc. (%) on RATS Arabic Levantine

Chan	SS	MS	MS-PM	Oracle
src	47.99	48.95	53.33	59.65
A	30.16	30.04	30.36	33.49
B	30.10	32.75	32.64	36.72
C	21.85	22.41	23.10	26.12
D	39.92	42.13	42.63	47.48
E	30.02	32.51	35.11	41.55
F	35.95	37.91	41.56	48.18
G	41.50	44.84	48.31	54.53
H	31.37	35.11	35.58	42.05

6. CONCLUSION

In this study we developed a subband hybrid (SBH) feature for multi-stream speech recognition. It employs two filterbanks of different bandwidth and spectral sampling rate for the representation of both long-term temporal modulation and short-term spectral modulation within a frequency subband. Experimental results indicate that the use of SBH feature significantly enhances the amount of speech information being extracted from individual subbands. Accordingly, the multi-stream system achieves a phoneme accuracy about 20–40% higher than the single-stream baseline on TIMIT database, with most of the gain in performance coming from the use of SBH feature. For narrow-band RATS Arabic Levantine the MS-PM system also significantly improves the phoneme accuracy for clean and most noisy channels. The gain deminishes when the speech sounds are corrupted in all subbands due to heavy noise or unknown distortion.

7. ACKNOWLEDGEMENT

The author would like to express his sincere thankfulness to Prof. Hynek Hermsnaky for insightful comments. The author also wants to thank Sriram Ganapathy, Samuel Thomas, G. S. V. S. Sivaram, Harish Sri Mallidi, Vijay Peddinti, Phani Nidadavolu, and other colleagues for sharing their work.

8. REFERENCES

[1] Allen, J. B. (1994), "How do humans process and recognize speech?", IEEE Trans. Speech and Audio, 2(4), 567–577.

[2] Athineos, M., Ellis, D. P. W. "Autoregressive modelling of temporal envelopes", IEEE Trans. Signal Processing. 55(11):5237–5245, 2007.

[3] Boothroyd, A., and Nittrouer, S. "Mathematical treatment of context effects in phoneme and word recognition." J. Acoust. Soc. Am. 84: 101, 1988.

[4] Bourlard, H. and Morgan, N., "Connectionist speech recognition: a hybrid approach", Springer 1994.

[5] Fletcher, H., and Galt, R. H. "The perception of speech and its relation to telephony." J. Acoust. Soc. Am. 22:89–151, 1950.

[6] Ganapathy, S., Thomas, S., and Hermansky, H., "Temporal envelope compensation for robust phoneme recognition using modulation spectrum.", J. Acoust. Soc. Am. 128(6):3769–3780, 2010.

[7] Hermansky, H., Variani, E., Peddinti, V. "Method for ASR performance prediction based on temporal characteristics of speech sounds", IEEE ICASSP 2013.

[8] Hermansky, H., "Speech recognition from spectral dynamics", Proc. Indian Academy of Sciences. 36(5):729–744, 2011.

[9] Kim, C. and Stern, R. M. "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition." IEEE Trans. Audio, Speech, and Language Processing, 2013.

[10] Li, F. and Jont B. Allen. "Multiband product rule and consonant identification." J. Acoust. Soc. Am. 126: 347, 2009.

[11] Li, F. and Hermansky, H. "Effect of filter bandwidth and spectral sampling rate on phoneme recognition", IEEE ICASSP 2013.

[12] Li, F., Mallidi, H., and Hermansky, H. "Phone recognition in critical bands using sub-band temporal modulations," Interspeech 2012

[13] McCourt, PM and Vaseghi, SV and Doherty, B., "Multi-resolution sub-band features and models for HMM-based phonetic modelling", Computer Speech & Language, 14(3):241–259, 2000.

[14] Ming, Ji, Peter Jancovic, and F. Jack Smith., "Robust speech recognition using probabilistic union models." IEEE Trans. Speech and Audio Processing, 10(6): 403-414, 2002.

[15] Sharma, S., "Multi-stream approach to robust speech recognition.", Ph.D. Thesis, Oregon Graduate Institute of Science and Technology, Portland. 1999.

[16] Tchorz, J., and Kollmeier, B. "A model of auditory perception as front end for automatic speech recognition." J. Acoust. Soc. Am. 106(4): 2040–2050, 1999.

[17] Variani, Ehsan and Li, Feipeng and Hermansky, Hynek, "Multi-stream recognitio of noisy speech with performance monitoring", Interspeech, 2013.