M-DELTA MEASURE FOR ACCURACY PREDICTION AND ITS APPLICATION TO MULTISTREAM-BASED UNSUPERVISED ADAPTATION

Tetsuji Ogawa[1], *Sri Harish Mallidi*[2], *Emmanuel Dupoux*[3], *Jordan Cohen*[4], *Naomi H. Feldman*[5], *Hynek Hermansky*[2]

[1]Waseda University, [2]Johns Hopkins University, [3]Ecole Normale Supérieure, [4]Spelamode, [5]University of Maryland

ABSTRACT

A new measure for predicting estimation accuracy is proposed and successfully applied to multistream-based unsupervised adaptation of ASR systems. The proposed measure is an extension of the M-measure, which predicts confidence in the output of a probability estimator by measuring divergences of probability estimates spaced at specific time intervals. In the present study, the M-measure was extended by considering latent phoneme information, resulting in improved reliability. Experimental comparisons carried out in a multistream-based ASR paradigm demonstrate that the extended M-measure yields a significant improvement over the original M-measure, especially in narrow-band noise conditions.

Index Terms— M-delta measure, performance monitor, multistream ASR

1. INTRODUCTION

When recognizing data at test time, automatic speech recognition (ASR) systems and other stochastic machines simply make their best guess on the basis of the data on which they were trained. Perfect learning could theoretically be achieved using infinitely large samples of data that covers all possible types of unexpected harmful variability that could be encountered during run-time of recognizers, but in practice such ideal learning is impossible. Creating ASR systems that adapt to environmental changes provides a way to address this fundamental weakness of machine learning.

Human listeners are able to estimate their confidence in decisions when perceiving degraded speech data that results in poor performance in ASR systems, even when the answer is not known *a priori* [1, 2]. Techniques for predicting the accuracy of an estimator based on the output of that estimator (e.g., estimates of phoneme posterior probabilities) have played an important role for unsupervised adaptation of ASR systems. For example, confusion networks [3] have been used to predict accuracy for each phoneme or word [4, 5]. Other measures that are computed over several seconds of speech (e.g., an utterance) can yield more reliable predictions of estimation accuracy.

Mean temporal distance (denoted "M-measure") [6], which evaluates averaged dissimilarities of probability estimates spaced in several time spans, is one such measure that has been shown to be effective in predicting estimation accuracy. This method makes use of the fact that vectors of phoneme probability estimates should be dissimilar between distant frames of speech, which are likely to belong to different phonemes. However, the M-measure does not explicitly consider more detailed information about the phonemic structure of speech.

The present study builds on the M-measure in two ways to develop improved confidence measures for use in multistream-based adaptation of recognizers that are robust against stream-specific noise. First, a new version of the M-measure is proposed that explicitly takes into account the probability that distant frames have different phoneme labels, providing a more accurate indicator of the estimator's ability to distinguish between phonemes. Next, these confidence measures are computed using probabilities of broad phoneme classes instead of standard phoneme probabilities. Using broad phoneme classes increases the coverage of phoneme classes in each sentence, and thus can potentially improve the reliability of sentence-by-sentence confidence estimation. The proposed techniques for confidence estimation are evaluated in a multistream-based adaptation paradigm [7], which is directly applicable to the current ANN-based ASR systems.

The paper is organized as follows. Section 2 gives an overview of relevant previous work on prediction of estimation accuracy and the multistream-based paradigm in ASR. Section 3 describes M-delta, our extension of the M-measure. Section 4 demonstrates that this M-delta measure is effective as an accuracy predictor and can be successfully applied to multistream-based adaptation of ASR systems, using both standard phoneme posteriors and broad phoneme classes. Section 5 concludes.

2. RELEVANT PREVIOUS WORK

2.1. Prediction of Estimation Accuracy

Several attempts have been made to predict estimation accuracy of ASR systems. We refer to these predictors as "performance monitors." Okawa et al. [8] used the entropy of the estimator outputs as a performance monitor. Ikbal et al. [9] and Kubo et al. [10] employed this measure for combining evidences from multiple ASR systems for noise robust ASR. Mesgarani et al. [11] and Badiezadegan et al. [12] computed the distance in autocorrelation of phoneme posterior probabilities between training and testing data. In addition, Variani and Hermansky [13] used Mahalanobis distance on logarithmic posterior space between training and testing data. The experimental results indicated that these criteria for performance monitor worked reasonably well but required a minimum of four seconds to obtain stable estimates of the probability distribution for posterior data. Ogawa et al. [14] demonstrated that likelihoods computed from Gaussian mixture model of the classifier outputs could be applied to predict estimation accuracy online (i.e., frame-by-frame). This

This work was supported in part by the National Science Foundation via award number IIA-0530118 and Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013.

criterion worked better than the aforementioned criteria even less than four seconds of speech segments. Another recently-proposed technique evaluates averaged dissimilarities of probability estimates spaced in several time apart, which has been named "M-measure" [6, 15]. This measure is simple but has been shown to be effective in predicting estimation accuracy [6]. The present work is an attempt to improve this measure and was extensively studied in 2014 Frederick Jelinek Memorial Workshop in Prague [1] [16].

2.2. Multistream-Based Unsupervised Adaptation of ASR

In multistream-based adaptation paradigm, reliable band-limited streams are chosen on the basis of the performance monitor and then fused. The first stage of parallel processing estimates the posterior probabilities of phonemes in band-limited streams. This is followed by a fusion stage that integrates the classification results from the band-limited streams on the basis of performance monitor. Sharma [17] proposed a prototype system of multistream ASR in which the full frequency was divided into seven bands to emulate the parallel processing that was hypothesized in human speech recognition, and to deal with selectively corrupted streams. All 127 non-empty combinations of these seven band-limited streams were formed and the second stage MLP classifier was trained for each of these 127 combinations. Our work is also based on multistream speech recognition but the full frequency was divided into five bands, yielding 31 processing streams.

The key to the success of multistream-based unsupervised adaptation of ASR systems lies in performance monitor, which predicts estimation accuracy of individual streams without requiring knowledge about correct answers. Several unsupervised techniques using the aforementioned measures have been proposed and investigated to select the least corrupted processing streams [11, 12, 13, 15, 18, 19].

3. EXTENSION OF M-MEASURE

An attempt is made to extend the M-measure. The original M-measure evaluates divergences in probability estimates across times without any consideration of phoneme contexts. In contrast, the extension of this measure, which is inspired by the segmentation algorithm proposed in [20], computes difference in divergences coming from the same phoneme and from different phonemes. In this section, the original M-measure is briefly explained and then the extended M-measure is described in detail.

3.1. M-measure

The original M-measure accumulates divergences between probability estimates spaced in several time-spans. It is defined as

$$\mathcal{M}(\Delta t) = \frac{1}{T-\Delta t}\sum_{t=\Delta t}^{T}\mathcal{D}(\mathbf{p}_{t-\Delta t}, \mathbf{p}_t), \tag{1}$$

where Δt denotes a time interval between phoneme posterior probabilities at $t-\Delta t$ and t, $\mathbf{p}_{t-\Delta t}$ and $\mathbf{p}_t$, and $\mathcal{D}(\mathbf{p},\mathbf{q})$ denotes the symmetric KL divergence between the posteriors,

$$\mathcal{D}(\mathbf{p},\mathbf{q}) = \sum_{k=0}^{K} p^{(k)} \log\frac{p^{(k)}}{q^{(k)}} + \sum_{k=0}^{K} q^{(k)} \log\frac{q^{(k)}}{p^{(k)}}, \tag{2}$$

[1] http://www.clsp.jhu.edu/workshops/archive/ws14-summer-workshop/groups/self-monitoring-asr/

where $p^{(k)}$ denotes the k-th element of a posterior vector $\mathbf{p} \in \mathbb{R}^K$. It has been observed that if an ASR system is developed using clean speech, this M-measure is higher for clean speech utterances (i.e., known data) and lower for noisy speech utterances (i.e., unknown data). In addition, as the signal-to-noise ratio (SNR) of noisy speech decreases, the M-measure tends to get lower. This means that the M-measure could be effective in determining whether the input data are known or unknown for the system. In multistream ASR, the stream (or system) with the highest M-measure can be selected as the most reliable stream (or system) [15].

The M-measures described in Eq. (1) are averaged over several time intervals Δt and the resulting is used as a confidence measure,

$$\mathcal{M} = \underset{\{\Delta t\}}{\text{mean}}[\mathcal{M}(\Delta t)], \tag{3}$$

where $\{\Delta t\}$ consists of 10, 15, 20, $\cdots$, 80 frames (15 intervals) in the present paper.

3.2. M-delta Measure

An extension of the M-measure, which is denoted "M-delta measure," computes the probability in each time span of two frames being an instance of the same phoneme. At test time, it estimates the M-measures for same vs. different phonemes by solving a redundant set of linear equations.

The original M-measure assumes that the distance between probability estimates in several time-spans should be large for known data (mainly for clean speech). However, this is not always accurate. If two posteriors are of the same phoneme class, the distance between them should be small, irrespective of time intervals. This means that the original M-measure ignores the effect of posterior pairs that are separated by large time intervals but belong to the same phoneme class. It accumulates symmetric KL divergence between posteriors without considering this kind of phoneme dependency.

We therefore introduce the idea of within-class and across-class M-measures, $\mathcal{M}^{\text{wc}}$ and $\mathcal{M}^{\text{ac}}$, to represent the accumulated KL-divergence computed from data pair of the same phoneme class and that from data pair of different classes, respectively. The new M-delta measure is defined using those within- and across-class M-measures as

$$\mathcal{M}delta = \mathcal{M}^{ac} - \mathcal{M}^{wc}. \tag{4}$$

Specifically, it is assumed that the M-measure can be decomposed into

$$\mathcal{M}(\Delta t) = p^{\text{wc}}(\Delta t)\cdot\mathcal{M}^{\text{wc}} + p^{\text{ac}}(\Delta t)\cdot\mathcal{M}^{\text{ac}} + \epsilon_{\Delta t}, \tag{5}$$

where $\mathcal{M}(\Delta t)$ denotes the original M-measure defined in Eq. (1), which is obtained for each utterance; $p^{\text{wc}}(\Delta t)$ and $p^{\text{ac}}(\Delta t)$ denote the prior probability of a pair of frames separated by Δt being instances of the same and different phonemes, respectively; and $\mathcal{M}^{\text{wc}}$ and $\mathcal{M}^{ac}$ denote the within-class and across-class M-measures being estimated for each utterance. The probabilities of frames belonging to the same vs. different phoneme classes at each time interval are obtained using the exact transcriptions of the training data.

The error term $\epsilon_{\Delta t}$ is included because Eq. (5) is an approximate representation of the M-measure. Although the prior probabilities computed from training data are reliably estimated, these probabilities vary across test utterances, because the variety of phonemes in a

test utterance is limited. To minimize the overall error of within-class and across-class M-measures, the redefined M-measure described in Eq. (5) can be written redundantly with several values of Δt. Assume that $\mathbf{y}$, $\mathbf{A}$, $\mathbf{x}$, and ϵ are given as follows:

$$\mathbf{y} = \left[\begin{array}{ccc} \mathcal{M}(\Delta t_1) & \cdots & \mathcal{M}(\Delta t_N) \end{array} \right]^{\mathrm{T}} \in \mathbb{R}^N \quad (6)$$

$$\mathbf{A} = \left[\begin{array}{cc} p^{\mathrm{wc}}(\Delta t_1) & p^{\mathrm{ac}}(\Delta t_1) \\ \cdots & \cdots \\ p^{\mathrm{wc}}(\Delta t_N) & p^{\mathrm{ac}}(\Delta t_N) \end{array} \right] \in \mathbb{R}^{N \times 2} \quad (7)$$

$$\mathbf{x} = \left[\begin{array}{cc} \mathcal{M}^{\mathrm{wc}} & \mathcal{M}^{\mathrm{ac}} \end{array} \right]^{\mathrm{T}} \in \mathbb{R}^2 \quad (8)$$

$$\epsilon = \left[\begin{array}{ccc} \epsilon_{t_1} & \cdots & \epsilon_{t_N} \end{array} \right]^{\mathrm{T}} \in \mathbb{R}^N \quad (9)$$

Then, Eq. (5) can be written as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \epsilon. \quad (10)$$

In this case, the within-class and across-class M-measures can be estimated as a least square solution as

$$\mathbf{x} = (\mathbf{A}^{\mathrm{T}}\mathbf{A})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}. \quad (11)$$

The experiments below used values of $(\Delta t_1, \Delta t_2, \cdots, \Delta t_N)$ = (1, 2, 3, 4, 5, 10, 15, 20, $\cdots$, 75, 80) and $N = 20$, which were determined from preliminary experiments.

4. MULTISTREAM-BASED UNSUPERVISED ADAPTATION OF ASR SYSTEM

The techniques for predicting estimation accuracy were evaluated as a performance monitor in multistream-based unsupervised adaptation of ANN speech recognition systems [7].

4.1. Multistream ASR System

The multistream-based adaptation paradigm used was introduced in [19]. The full frequency of the speech signal is divided into five band-limited streams, each of which covers about three barks along auditory frequency. Then, the processing streams are formed for all non-empty combinations of five band-limited streams, yielding 31 processing streams. The most reliable processing stream was selected using performance monitors and the posterior probabilities from the ANN for that stream were used for obtaining final recognition results. This adaptation paradigm can yield advantages in band-limited noise corruption by utilizing a stream that does not contain the corrupted band.

In each band-limited stream, temporal modulation information was extracted from 250 ms temporal envelopes using frequency domain linear prediction (FDLP) analysis [21]. An ANN-based probability estimator was trained for each band-limited stream with inputs as the corresponding FDLP features and triphone states as targets. The ANNs have four hidden layers of 1024 units, input layer of 576 nodes, and 1951 output units. This band-limited ANNs were used to yield 39-dimensional phoneme posterior probabilities. In the latter stage, ANN-based probability estimators were developed for 31 processing streams. The features were obtained by stacking the phoneme posterior probabilities from the band-limited ANNs.

4.2. Confidence Measures for Performance Monitor

Experimental comparisons were conducted for three measures:

- **M**: original M-measure
- **Mdelta**: M-delta measure with $\mathcal{M}^{\mathrm{wc}}$ and $\mathcal{M}^{\mathrm{ac}}$
- **Entropy**: accumulated negative entropy of phoneme posterior probabilities

In the present experiments, these measures were computed based on single sentence to predict accuracy for that sentence.

Table 1. *Types and SNRs of noise used.*

item	noise type	SNR
clean		
sub15	subway	15
bab15	babble	15
fac10	factory	10
res10	restaurant	10
exh5	exhibition hall	5
str5	street	5
car5	car	5
exh0_b2	exhibition hall (band 2 corrupted)	0
exh0_b4	exhibition hall (band 4 corrupted)	0

4.3. Speech Materials

All models were trained on 3696 clean speech utterances from TIMIT training data set and the evaluation was conducted using 400 speech utterances from the TIMIT development set under several types of noise. The types and SNRs of noise are listed in Table 1. Note that in principle, the multistream-based adaptation paradigm enables an ASR system to be more robust against stream-specific noise, such as the exh0_b2 and exh0_b4 noises listed in Table 1.

4.4. Experimental Results

Evaluation criteria are

- how well the scores from the performance monitor correlate with actual recognition accuracies
- phoneme error rate of an ASR system with multistream-based unsupervised adaptation

Note that the aim of multistream-based unsupervised adaptation is selecting the most reliable processing stream from 31 streams for each sentence. Ideally, the confidence measures obtained from 31 processing streams should correlate highly with the corresponding phoneme error rates for each utterance. Correlations with the actual phoneme accuracy were therefore calculated for each utterance individually across 31 processing streams and then averaged over the 400 utterances of the TIMIT development set. Phoneme error rates were obtained from the processing stream estimated by the performance monitor for each utterance, averaged over 400 utterances.

4.4.1. Evaluation by correlations with phoneme accuracy

Figure 1 shows the correlation between the confidence measure and actual phoneme accuracy for several types of noise. This figure shows that the M-delta measure yielded significant improvement over the existing measures, such as the negative entropy and original M-measure, in the narrow-band noise conditions, i.e., exh0_b2 and exh0_b4, while it gave similar performance to the original M-measure and did not yield an advantage over the entropy under the broad-band noise corruptions.

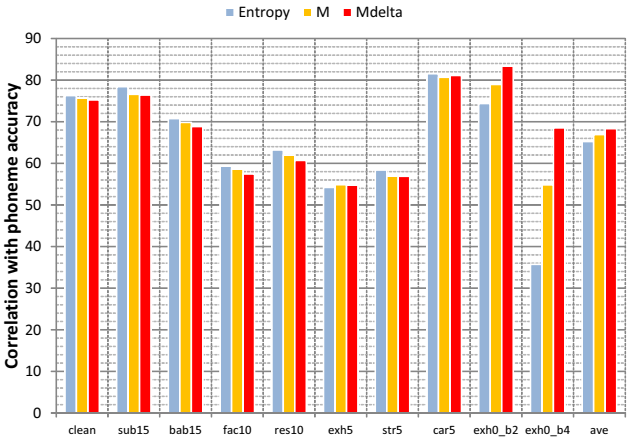

Fig. 1. Correlations with phoneme accuracy in multistream-based adaptation for several types of noise. Bars for "ave" express correlations averaged over ten conditions.

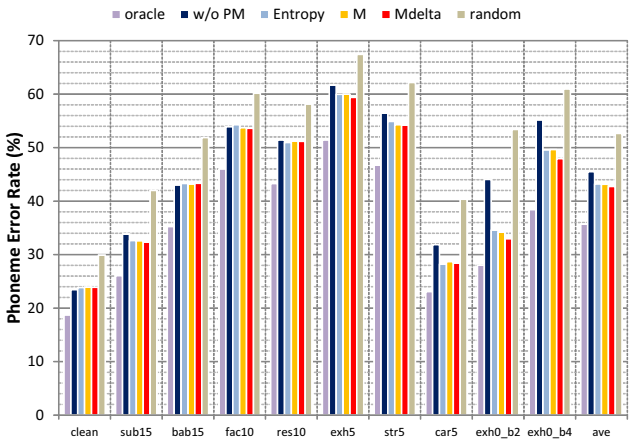

Fig. 2. Phoneme error rates obtained by multistream-based unsupervised adaptation for several types of noise. Bars for "ave" express phoneme error rates averaged over ten conditions.

4.4.2. Evaluation by selecting stream in multistream ASR

The stream that yielded the highest confidence estimate was selected from 31 recognizers, using accuracy prediction. The resulting phoneme error rate was calculated from the recognizer outputs of the selected stream. Experimental comparisons were as follows:

- **Oracle**: selecting the stream with the best error rate by hand
- **w/o PM**: including all individual band-limited streams [22]
- **random**: selecting a stream at random
- **w/ PM**: selecting a stream with the performance monitor

The negative entropy, original M-measure, and M-delta measure were used for the systems **w/ PM**.

Figure 2 shows phoneme error rates for several types of noise. This figure demonstrates that multistream-based unsupervised adaptation with the performance monitor (**w/ PM**) can reduce phoneme errors from the system without the performance monitor (**w/o PM**) and that based on random selection of a processing stream (**random**). Specifically, the M-delta measure yielded a small but consistent advantage in broad-band noise corruptions and more significant gains in the narrow-band noise conditions.

4.4.3. Use of broad phoneme class probability estimator

The techniques for predicting estimation accuracy were modified by using estimates of probabilities of broad phoneme classes instead of estimates of standard phoneme probabilities. Using broad phoneme classes can deteriorate accuracy of M-measures but improve their reliability by increasing a coverage of classes. The broad phoneme classes used are defined in [23], i.e., plosives, fricatives, nasals, semi-vowels, vowels, diphthongs, and silence.

Figure 3 shows the correlation with phoneme accuracy for several types of noise when broad phoneme classes are used. Using the broad phoneme class probability estimator yielded significant improvement, irrespective of the measures, in the narrow-band noise conditions, but did not help in broad-band noise corruption.

The M-delta measure and use of broad phoneme class probabilities were extensively studied in 2014 Frederick Jelinek Memorial Workshop in Prague. During this workshop, the proposed techniques were compared to traditional confidence measures such as acoustic and language model likelihood along with lattice and posterior probabilities in confusion sets accumulated over single sentence. This comparison demonstrated that the M-delta measure yielded best results in the narrow-band noise conditions [16].

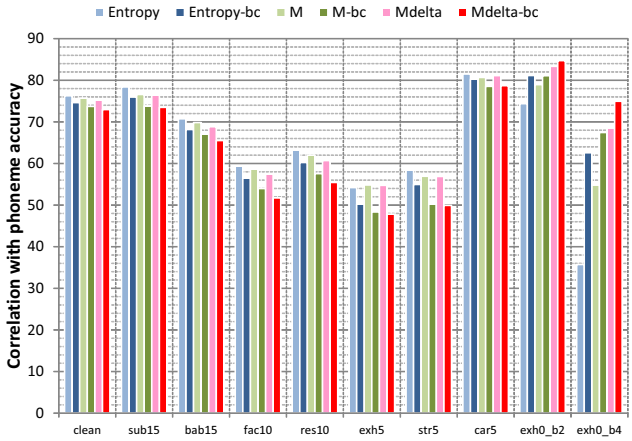

Fig. 3. Correlation with phoneme accuracy in multistream-based adaptation for several types of noise when using broad phoneme class probabilities (Entropy-bc, M-bc, and Mdelta-bc) and standard phoneme class probabilities (Entropy, M, and Mdelta).

5. CONCLUSION

The M-measure was extended and successfully applied to multistream-based unsupervised adaptation in ASR. The within-class and across-class M-measures were introduced to consider phoneme class information that was ignored in the original M-measure and obtained by solving redundant set of equations. This extension (M-delta measure) yielded significant gains from the original M-measure, especially under the narrow-band noise. By reducing the influence of phoneme contexts on the confidence measures, using broad phoneme class probabilities instead of standard phoneme probabilities in the performance predictor yielded further improvement. Both of these cases suggest that taking into account what is known about the structure of phonemes in speech can lead to the creation better adaptive speech technologies.

6. REFERENCES

[1] M. K. Sheffers and M. G. H. Coles, "Performance monitoring in confusing word: Error brain activity, judgments of response accuracy, and types of errors," *J. Exp. Psych.*, vol. 26, no. 1, pp. 141–151, 2000.

[2] J. D. Smith and D. A. Wahsburn, "Uncertainty monitoring and metacognition by animals," *Current Directions In Psychological Science*, vol. 14, no. 1, pp. 19–24, 2005.

[3] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion network," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[4] J. Ogata and Y. Ariki, "Unsupervised acoustic model adaptation based on phoneme error minimization," in *Proc. ICSLP*, Sept. 2002, pp. 1429–1432.

[5] H. Xu, D. Povey, L. Mangu, and J. Zhu, "Minimum bayes risk decoding and system combination based on a recursion for edit distance," *Computer Speech and Language*, vol. 25, no. 4, pp. 802–828, 2011.

[6] H. Hermansky, E. Variani, and V. Peddinti, "Mean temporal distance: predicting ASR error from temporal properties of speech signal," in *Proc. ICASSP*, May 2013, pp. 7423–7426.

[7] H. Hermansky, "Multistream recognition of speech: dealing with unknown unknowns," *Proc. IEEE*, vol. 101, no. 5, pp. 1076–1088, 2013.

[8] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proc. ICASSP*, 1998, vol. 2.

[9] S. Ikbal, H. Misra, S. Sivadas, H. Hermansky, and H. Bourlard, "Entropy based combination of tandem representations for noise robust ASR," in *Proc. INTERSPEECH*, Oct. 2004.

[10] Y. Kubo, O. Okawa, A. Kurematsu, and K. Shirai, "Noisy speech recognition using temporal AM-FM combination," in *Proc. ICASSP*, April 2008, pp. 4709–4712.

[11] N. Mesgarani, S. Thomas, and H. Hermansky, "Adaptive stream fusion in multistream recognition of speech," in *Proc. Interspeech*, Aug. 2011, pp. 2329–2332.

[12] S. Badiezadegan and R. Rose, "A performance monitoring approach to fusing enhanced spectrogram channels in robust speech recognition," in *Proc. Interspeech*, Aug. 2011, pp. 4780–4783.

[13] E. Variani and H. Hermansky, "Estimating classifier performance in unknown noise," in *Proc. Interspeech*, Sept. 2012.

[14] T. Ogawa, F. Li, and H. Hermansky, "Stream selection and integration in multistream ASR using GMM-based performance monitoring," in *Proc. INTERSPEECH*, Aug. 2013, pp. 3332–3336.

[15] E. Variani, F. Li, and H. Hermansky, "Multi-stream recognition of noisy speech with performance monitoring," in *Proc. Interspeech*, Aug. 2013.

[16] H. Hermansky et al., "Towards machines that know when they do not know: Summary work done at 2014 Frederick Jelinek Memorial Workshop in Prague," (*submitted*).

[17] S. Sharma, "Multi-stream approach to robust speech recognition," *Ph. D Thesis, Oregon graduate institute of science and technology, Portland*, 1999.

[18] N. Mesgarani, S. Thomas, and H. Hermansky, "A multistream multiresolution framework for phoneme recognition," in *Proc. Interspeech*, Sept. 2010, pp. 318–321.

[19] F. Li, "Subband hybrid feature for multi-stream speech recognition," in *Proc. ICASSP*, 2014, pp. 2484–2488.

[20] J. Cohen, "Segmenting speech using dynamic programming," *J. Acoust. Soc. Amer*, vol. 69, no. 5, pp. 1430–1438, 1981.

[21] S. Ganapathy, S. Thomas, and H. Hermansky, "Temporal envelope compensation for robust phoneme recognition using modulation spectrum," *J. Acoust. Soc. Amer.*, vol. 128, no. 6, pp. 2769–3780, 2010.

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

[23] T. J. Reynolds and C. A. Antoniou, "Experiments in speech recognition using a modular MLP architecture for acoustic modeling," *Information Sciences*, vol. 156, pp. 39–54, 2003.