A new efficient measure for accuracy prediction and its application to multistream-based unsupervised adaptation

Tetsuji Ogawa*, Sri Harish Mallidi†, Emmanuel Dupoux‡, Jordan Cohen§
Naomi H. Feldman¶ and Hynek Hermansky†
*Waseda University, †Johns Hopkins University, ‡Ecole des Hautes Etudes Sciences Sociale
§Spelamode, ¶University of Maryland

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

I. Introduction

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 when attempting to recognize data during test time. Perfect learning could theoretically be achieved using infinitely large samples of data that cover all possible types of unexpected harmful variables that could be encountered during run-time of the recognizers, but in practice such an ideal learning is impossible. Creating ASR systems that adapt to the environmental changes provides a way to address this fundamental machine learning weakness.

Human listeners are able to estimate their confidence in their decisions when perceiving degraded speech data, whereas ASR systems would poorly perform, even when the answer is not known *a priori* [1], [2]. Techniques for predicting the accuracy of an estimator based on its output (e.g., estimates of phoneme posterior probabilities) have played an important role in the unsupervised adaptation of ASR systems. For example, confusion networks [3] have been used to predict the accuracy of 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 the estimation accuracy.

The mean temporal distance (denoted as the "M-measure") [6], which evaluates the averaged dissimilarities in the probability estimates spaced over several time spans, is one such measure that has been proven to be effective in predicting the estimation accuracy. This method makes use of the fact that the vectors of the phoneme probability estimates should be dissimilar between the 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 to develop improved confidence measures for use in the multistream-based adaptation of recognizers that are robust against noise. 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. The proposed technique for the confidence estimation is evaluated using a multistream-based adaptation paradigm [7], which is directly applicable to the current DNN-based ASR systems.

The paper is organized as follows. Section II gives an overview of the relevant previous work on the prediction of the estimation accuracy and the multistream-based paradigm in ASR. Section III describes M-delta, our extension of the M-measure. Sections IV and V demonstrate that this M-delta measure is effective for use as an accuracy predictor and can be successfully applied to two types of multistream-based adaptation of ASR systems. Section VI concludes this paper.

II. Relevant previous work

A. Prediction of estimation accuracy

Several attempts have been made to predict the 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] used this measure for combining results from multiple ASR systems for noise robust ASR. Mesgarani et al. [11] and Badiezadegan et al. [12] computed the distance in the autocorrelation of the phoneme posterior probabilities between training and testing data. In addition, Variani and Hermansky [13] used the Mahalanobis distance on the logarithmic posterior space between training and testing data. The experimental results indicated that these criteria as a 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 the likelihoods computed from the Gaussian mixture model of the classifier outputs could be

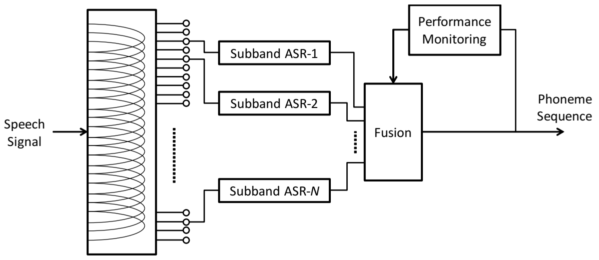

Fig. 1. *Schematic diagram of multistream-based adaptation paradigm. Uncorrupted streams are selected from all band-limited streams on basis of performance monitor and then fused.*

applied for predicting the estimation accuracy frame-by-frame. This criterion worked better than the aforementioned criteria with even less than four seconds of speech. Another recently-proposed technique evaluates the averaged dissimilarities in the probability estimates spaced several time apart, which has been named the "M-measure" [6], [15]. This measure is simple but has proven to be effective in predicting the 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 [16], [17].

B. Multistream-based unsupervised adaptation of ASR

Figure 1 depicts a schematic diagram of multistream-based adaptation paradigm, in which reliable band-limited streams are chosen on the basis of the performance monitor and then fused. In the multistream-based adaptation paradigm, reliable band-limited streams are chosen on the basis of the performance monitor and then fused [7]. The first stage of the parallel processing estimates the posterior probabilities of phonemes in the band-limited streams. This is followed by a fusion stage that integrates the classification results from the band-limited streams on the basis of the performance monitor. Sharma [18] proposed a prototype multistream ASR system in which the full frequency was divided into seven bands to emulate the parallel processing that was hypothesized in human speech recognition, and to selectively deal with 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 ASR, but the full frequency is divided into five bands.

The key to the success of the multistream-based unsupervised adaptation of ASR systems lies in the performance monitor, which predicts the estimation accuracies of individual streams without requiring any knowledge about the correct answers. Several unsupervised techniques using the aforementioned measures have been proposed and investigated for selecting the least corrupted streams [11], [12], [13], [15], [19], [20].

Audio-visual ASR is an alternative multistream approach to creating noise robust ASR systems [21]. This is not the focus in the present study, but unsupervised adaptation based on the performance monitoring is also applicable to this approach.

III. Extension of M-measure

An attempt has been made to extend the M-measure. The original M-measure evaluates the divergences in probability estimates across times without any consideration of the phoneme contexts. The extension of this measure, which was inspired by the segmentation algorithm proposed in [22], computes the difference in divergences coming from the same phoneme as well as different phonemes. This section briefly explains the original M-measure and describes the extended M-measure in detail.

A. M-measure

The M-measure accumulates the divergences of probability estimates spaced over 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 the time interval between the 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,

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

where $p^{(k)}$ denotes the k-th element of a posterior vector $\mathbf{p} \in \mathbb{R}^K$. It has been found 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 SNR of noisy speech decreases, the M-measure lowers. This means that the M-measure could be effective in determining whether the input data are known or unknown for a 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 in Eq. (1) are averaged over several time intervals Δt and the result is used as the 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).

B. M-delta measure

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

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

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

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

We assume that the M-measure can be decomposed into

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

where $\mathcal{M}(\Delta t)$ denotes the original M-measure defined using Eq. (1), which is determined for each utterance; $p^{\mathrm{wc}}(\Delta t)$ and $p^{\mathrm{ac}}(\Delta t)$ denote the probability of a pair of frames separated by Δt being instances from the same and different phonemes, respectively; and $\mathcal{M}^{\mathrm{wc}}$ and $\mathcal{M}^{\mathrm{ac}}$, the within- and across-class M-measures being estimated for each utterance. $p^{\mathrm{wc}}(\Delta t)$ and $p^{\mathrm{ac}}(\Delta t)$ are determined from the training data transcriptions.

The error term $\epsilon_{\Delta t}$ is included because Eq. (5) is an approximate representation of the M-measure. Although $p^{\mathrm{wc}}(\Delta t)$ and $p^{\mathrm{ac}}(\Delta t)$, which are computed from the training data, are reliably estimated, these probabilities actually differ from those computed from the test utterances, because the variety of phonemes in a test utterance is limited. The redefined M-measure described using Eq. (5) can be written redundantly with several Δt values to minimize the overall error of the within- and across-class M-measures. Assume that $\mathbf{y}$, $\mathbf{A}$, $\mathbf{x}$, and ϵ are given as

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

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

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

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

Then, Eq. (5) can be written as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \epsilon. \tag{10}$$

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

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

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

IV. Multistream-based unsupervised adaptation of ASR system

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

TABLE I
Types and SNRs of noise used.

item	noise type	SNR [dB]
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

A. Multistream ASR system based on band-limited streams

The multistream-based adaptation paradigm used was introduced in [20]. The full frequency of the speech signal is divided into five band-limited streams, each of which covers about three barks along the auditory frequency. Then, the processing streams are formed for all the non-empty combinations of the five band-limited streams, yielding 31 processing streams. The most reliable processing stream was selected using performance monitors and the posterior probabilities from the DNN for that stream were used for determining the final recognition results. This adaptation paradigm yields advantages in terms of the band-limited noise corruption by using a stream that does not contain the corrupted band.

The temporal modulation information in each band-limited stream was extracted from 250ms temporal envelopes using frequency domain linear prediction (FDLP) analysis [23]. A DNN-based probability estimator was trained for each band-limited stream with inputs as the FDLP features and triphone states used as the targets. The DNNs have four hidden layers of 1024 units, an input layer of 576 nodes, and 1951 output units. These band-limited DNNs were used to yield 39-dimensional phoneme posterior probabilities. The DNN-based probability estimators were developed for the 31 processing streams in the latter stage. The features were determined by stacking the phoneme posterior probabilities from the band-limited DNNs.

B. Confidence measures for performance monitor

Experimental comparisons were conducted for three measures:

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

These measures were computed based on a single sentence to predict the accuracy for that sentence.

C. Speech materials

All the models described in IV-A and the probabilities p^{wc} and p^{ac} in Eq. (5) were trained on 3696 clean speech utterances from the TIMIT training 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 the noise are listed in Table I. There were 61

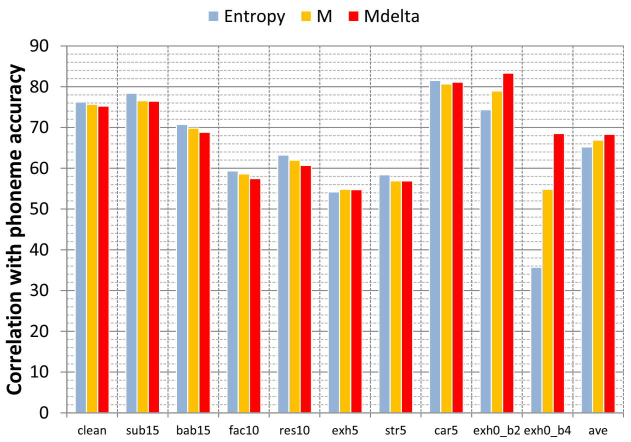

Fig. 2. Correlations with phoneme accuracy in multistream-based adaptation for several types of noise. “ave” bars represent correlations averaged over ten conditions.

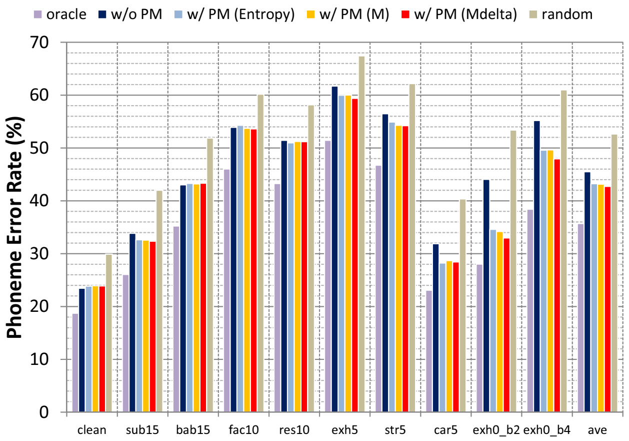

Fig. 3. Phoneme error rates determined using multistream-based unsupervised adaptation for several types of noise. “ave” bars represent error rates averaged over ten conditions.

phonemes in the TIMIT transcriptions that were compacted into 48 phonemes for training and 39 phonemes for evaluation, as proposed in [24]. 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.

D. Experimental results

The evaluation criteria were:

- How well the scores from the performance monitor correlated with the actual recognition accuracies
- The phoneme error rate of an ASR system with multistream-based unsupervised adaptation.

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

1) Evaluation by correlations with phoneme accuracy: Figure 2 shows the correlation between the confidence measure and the actual phoneme accuracy for several types of noise. This figure shows that the M-delta measure yielded a significant improvement over the existing measures, such as the negative entropy and original M-measure, under the narrow-band noise conditions, i.e., exh0_b2 and exh0_b4, while it yielded similar results to those for the original M-measure and did not yield an advantage over the entropy in the broad-band noise corruptions.

2) Evaluation by selecting stream in multistream ASR: The stream that yielded the highest confidence was selected from the 31 recognizers, using the accuracy prediction. The phoneme error rate was calculated from the recognizer outputs of the selected stream. The comparisons were as follows:

- **Oracle**: selecting the stream with the best error rate by hand
- **w/o PM**: including all the individual band-limited streams [25]
- **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 3 shows the phoneme error rates for several types of noise. This figure proves that the multistream-based unsupervised adaptation with the performance monitor (**w/ PM**) can reduce the amount of phoneme errors from a system without the performance monitor (**w/o PM**) and that based on the random selection of a processing stream (**random**). In particular, the M-delta measure yielded a small but consistent advantage in the broad-band noise corruptions and more significant gains under the narrow-band noise conditions.

3) Use of broad phoneme class probability estimator: The techniques for predicting the estimation accuracy were modified using the estimates of the probabilities of broad phoneme classes instead of the estimates of standard phoneme probabilities. Using the broad phoneme classes can deteriorate the accuracy of the M-measures but improve their reliability by increasing the coverage of the classes. The seven broad phoneme classes used were defined in [26], i.e., plosives, fricatives, nasals, semi-vowels, vowels, diphthongs, and silence. The posterior probabilities $\mathbf{p}_t$ for computing the negative entropy and M-measure are determined by merging the phoneme posteriors corresponding to a broad phoneme class. Note that such broad phoneme classes are used only for computing the confidence measures (i.e., stream selection) and the posterior probabilities for the triphone states are calculated during the recognition of the selected stream.

Figure 4 shows the correlation with the phoneme accuracy

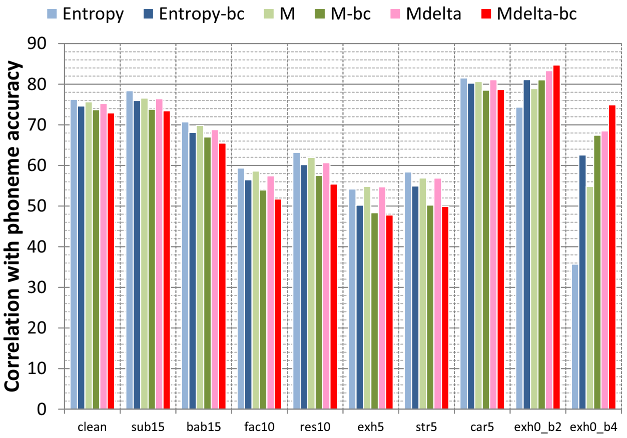

Fig. 4. Correlation with phoneme accuracy in multistream-based adaptation for several types of noise when using broad phoneme class (Entropy-bc, M-bc, and Mdelta-bc) and standard phoneme class probabilities (Entropy, M, and Mdelta).

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, under the narrow-band noise conditions, but did not help for the 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 the traditional confidence measures, such as the acoustic and language model likelihood along with lattice and posterior probabilities in confusion sets accumulated for a single sentence. This comparison demonstrated that the M-delta measure yielded the best results under the narrow-band noise conditions [17].

V. Multistream ASR based on selection of noise specific streams

In the experiments described in Section IV, the DNNs for all possible combinations of band-limited streams were trained with only the clean speech, and the developed multistream framework was robust against the band-limited noise corruption. In this section, another type of multistream-based unsupervised adaptation is discussed. In each stream, the DNN is trained on a specific noise condition. This results in a multistream framework where each stream performs well on a specific noise condition. For a given test utterance, selecting posterior estimates from the stream having the most similar acoustic property, results in the lowest error rate.

A. Experimental setup

We used 3696 utterances from the TIMIT training set and 400 utterances from the TIMIT development set for the purpose of testing. Ten types of original training set are created by corrupting the clean training speech with nine types of additive noise, at various signal-to-noise ratios (SNRs) ranging from 0 dB to 20 dB. We used babble, buccaneer1, buccaneer2, car, destroyerops, exhall, f16, factory1, and factory2 noises from NOISEX database.

The original clean training set and nine noisy training sets are combined to create a multi-condition training set, the amount of which is ten times as much as other sets. The eleven types (one clean, nine noisy, and one multi-condition) of training sets are used to train eleven different DNNs, where ten of them are trained on a specific acoustic condition, and one DNN is trained on multi-condition data. The DNNs used have a depth of six hidden layers, and each hidden layer consist of 1024 sigmoidal units. We used 40 dimensional Mel filter-bank energy features. The DNNs are pre-trained using RBM [27] and fine-tuned using the cross-entropy cost function. The targets used for fine-tuning are context dependent triphone states, generated using a GMM/HMM system.

Similar to the training set, we corrupted the development set with the nine types of noise at SNRs of 0, 5, 10, 15 and 20 dB. The whole development set (clean and noisy versions) is referred to as the test set from here on.

B. Experimental result

Table II shows the results of the test set in various streams. For the purpose of showing the upper limit of performance, the **oracle** selection technique is defined as selecting the stream that has the most similar acoustic condition of given test data. In the present study, we used two types of oracle stream selection techniques as follows:

- **Utterance oracle**: We select a stream with the lowest error rate for each utterance by hand.
- **Matched condition**: We select a stream trained on the same noise for a test utterance.

We can infer that error rates of the condition-level oracle streams (i.e., Matched condition) are always less than those of individual streams (i.e., clean, car, babble, and so on). In addition, the utterance-level oracle streams performs better than the condition-level oracle streams.

Uncertainty measures for stream selection are as follows:

- **Entropy**: Stream selection based on entropy minimization
- **M**: Stream selection based on M value maximization
- **Mdelta**: Stream selection based on M-delta maximization

Table II shows that the entropy of posterior probability, obtained at the output of DNN is erroneous. The M measure performs better than the entropy, which suggests measures that look at temporal dynamics of posteriors are better than those looking at a single frame. The M-delta measure yields the improvement over the M measure and multi-condition training. In addition, integration of two best streams selected by the M-delta measure (Mdelta-top2), in which the geometric mean of two DNN posteriors is used for decoding, matches with the condition-level oracle stream. This results show that the M-delta measure successfully selects condition specific streams.

VI. Conclusion

The M-measure was extended and successfully applied to the multistream-based unsupervised adaptation in ASR. The

TABLE II

PHONEME ERROR RATES (%) FROM STREAM-SELECTION SYSTEM USING UNCERTAINTY MEASURES AND INDIVIDUAL SYSTEMS.

Train \Test	clean	bab	buc1	buc2	car	des	exh	f16	fac1	fac2	Ave
clean	20.7	59.2	65.7	64.9	34.2	59.5	57.4	62.9	62.0	53.3	54.0
babble	29.2	35.6	47.0	49.9	32.0	43.6	37.2	41.5	42.8	32.4	39.1
bucc1	31.9	54.2	35.6	43.6	40.0	52.6	53.0	40.8	49.1	40.6	44.1
bucc2	35.8	58.7	43.8	35.2	44.4	51.9	56.0	44.8	51.3	46.8	46.9
car	23.7	58.0	64.8	64.2	22.7	55.9	54.2	62.7	60.5	48.6	51.5
destroyerops	28.4	47.9	44.1	43.3	31.7	33.7	45.5	42.1	44.0	36.4	39.7
exhall	29.5	40.3	46.1	48.4	31.7	42.1	33.7	42.0	43.0	32.5	38.9
f16	29.8	49.2	40.8	44.7	37.0	48.7	48.1	33.2	46.6	36.4	41.5
factory1	29.8	46.1	39.1	40.8	33.9	43.6	44.0	37.7	36.5	32.9	38.4
factory2	27.0	44.7	43.7	48.0	29.1	45.3	42.8	41.0	43.3	29.3	39.4
Multi-condition	22.8	36.8	38.7	39.6	25.0	34.8	34.3	36.2	36.3	28.9	33.3
Matched condition	20.7	35.6	35.6	35.2	22.7	33.7	33.7	33.2	36.5	29.3	31.6
Utterance oracle	17.6	31.8	30.9	31.5	20.0	30.0	29.7	29.1	31.7	24.4	27.7
Entropy	22.5	38.0	39.8	43.2	24.9	36.5	35.2	36.6	37.7	29.2	34.4
M	22.8	39.7	34.4	36.2	25.0	35.4	36.5	32.5	39.5	29.5	33.2
Mdelta	22.8	38.5	34.4	36.2	25.0	33.3	35.5	32.5	38.7	29.1	32.6
Mdelta-top2	20.1	37.6	34.2	35.3	21.6	33.6	34.7	32.5	38.5	28.3	31.6

within- and across-class M-measures were introduced to take the phoneme class information that was ignored in the original M-measure into consideration and determined by solving a redundant set of equations. This extension (M-delta measure) yielded significant gains over the original M-measure, especially when there was narrow-band noise, in selection of band-limited streams trained on clean speech. The improvement was made also in selection of streams formed on specific noise conditions. Both of these cases suggest that taking into account what is known about the structure of the phonemes in speech can lead to the creation of better adaptive speech technologies.

ACKNOWLEDGMENT

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

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*, vol. 2, 1998.

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

[17] H. Hermansky, L. Burget, J. Cohen, E. Dupoux, N. Feldman, J. Godfrey, S. Khudanpur, M. Maciejewski, S. Mallidi, A. Menon, T. Ogawa, V. Peddinti, R. Rose, R. Stern, M. Wiesner, and K. Vesely, "Towards machines that know when they do not know: Summary work done at 2014 Frederick Jelinek Memorial Workshop in Prague," in *Proc. ICASSP*, April 2015, pp. 5009–5013.

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

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

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

[21] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Seniro, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sept. 2003.

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

[23] 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.

[24] K. Lee and H. Hon, "Speaker-independent phone recognition using hidden markov models," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 37, no. 11, pp. 1642–1648, Nov. 1989.

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

[26] 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.

[27] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, July 2006.