Stream Selection and Integration in Multistream ASR Using GMM-Based Performance Monitoring

Tetsuji Ogawa[1,2]*, Feipeng Li*[2]*, Hynek Hermansky*[2,3]

[1]Dept. of Computer Science, Waseda University,
[2]Center for Language and Speech Processing, Johns Hopkins University,
[3]Human Language Technology Center of Excellence, Johns Hopkins University

Abstract

A moderately deep and rather wide artificial neural net is applied in phoneme recognition of noisy speech. The net is formed by first estimating posterior probabilities of phonemes in 21 band-limited streams covering the whole speech spectrum. These 21 band-limited streams are subdivided into three seven band-limited stream subsets, by differently sub-sampling the original 21 band-limited streams. In the second processing stage, all non-empty combinations of seven band-limited streams from each subset are formed as inputs to 127 artificial neural nets that are again trained to yield phoneme posteriors. In this way, 127 $\times$ 3 = 381 processing streams are formed. A novel technique for finding the best combination of the resulting 381 parallel processing streams, which uses the likelihood of a single-state Gaussian mixture model of the final classifier output is applied to selecting the most efficient streams. The technique is efficient in phoneme recognition of speech that is corrupted by realistic additive noise.

Index Terms: multistream speech recognition, multilayer perceptron, performance monitoring, Gaussian mixture model

1. Introduction

Human listeners are able to estimate their confidence in decisions even when the answer is not known *a priori* for such degraded speech data that automatic speech recognition systems cannot achieve good performance [1, 2]. It would be very useful in many engineering applications to be able to emulate such abilities. We seek support for a hypothesis where the frequency selective human cochlea creates multiple processing frequency channels that serve as independent channels for speech communication. The corruption of any one channel has little impact on the performance of the overall system. This hypothesis is supported by work by Fletcher and his colleagues who investigated the contribution of different frequency bands to human speech recognition [3]. They observed that the average phoneme error rate of full-band speech was equal to the product of error from individual bands, $e = e_1 e_2 \cdots e_N$, where $e_i, i = 1, \cdots, N$ denotes the average error rate when only frequency band i is audible. In other words, the human auditory system can recognize speech with few problems as long as at least one of the channels fields reliable recognition result. In contrast, typical ASR systems treat the full frequency band as one channel.

We consider a recognizer that consists of many different parallel processing streams. When encountering unexpected signal distortions for which the system is not trained, some streams in such multistream recognizers might perform better than others. Being able to adaptively select only the best processing streams for the further processing could make the recognizer degrade more gradually than it would without such feedback. The topic in our ongoing research efforts is the selection of reliable streams and the fusion of information from the selected streams. The current paper contributes to this effort.

The current paper describes an artificial neural network (ANN) based phoneme recognizer that is aimed at attaining such abilities. Twenty one band-limited streams in the first stage of processing, each using evidence from only part of the available speech spectrum, are designed to estimate the posterior probabilities of 40 phonemes. A very wide ANN is formed in the second processing stage by subdividing the original 21 band-limited streams into three different seven stream subsets, forming all 127 non-empty combinations of each of seven band-limited stream subsets, yielding 127 $\times$ 3 = 381 processing streams. A performance monitoring module at the output of each of these 381 parallel processing streams selects the best processing streams for a given situation. Posterior estimates from the best processing streams are then used to derive the final posterior probability vector that is used in a Viterbi search for the best phoneme sequence in a hidden Markov model - artificial neural network (HMM-ANN) hybrid recognizer [4].

We demonstrate that the proposed system can reduce the effect of unreliable streams that are corrupted by noise. The results obtained from the present study could be useful in developing a multistream speech recognition system that is robust against non-stationary noise.

The rest of the present paper is organized as follows. Section 2 describes relevant prior work. Section 3 briefly reviews the subband speech recognition system and fusion of multiple streams. Section 4 presents the online selection and integration of reliable processing streams. Section 5 describes our experimental evaluation of the proposed method in terms of its robustness against noise and efficiency. We present our conclusions in Section 6.

2. Relevant Work in Multistream ASR

Li et al. [5] recently developed a 21 band-limited stream system, which is illustrated in Fig. 1. This system successfully fused all band-limited streams into a single processing stream using a multilayer perceptron (MLP) ANN. However, when encountering unexpected distortions, the accuracy of such single stream ASR systems is affected by unreliable (i.e., noise corrupted) band-limited streams.

Sharma et al. [6] proposed a prototype system of multi-

This work was supported in parts by the DARPA RATS project D10PC0015, IARPA BABEL project W911NF12-C-0013, and by the Johns Hopkins Center of Excellence in Human Language Technologies.

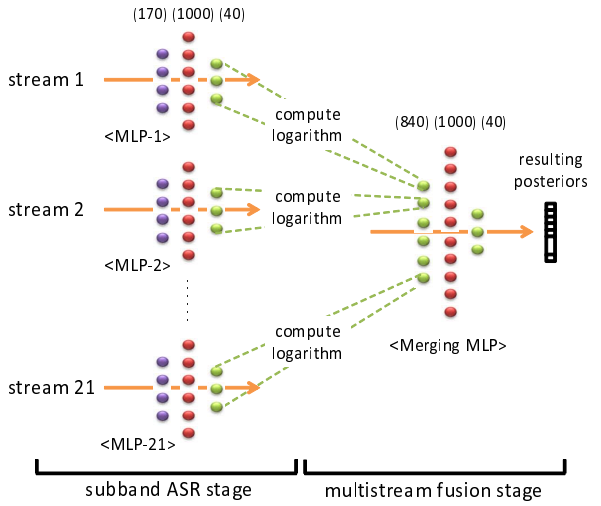

Figure 1: *Conceptual image of phoneme recognition system using merging MLP for fusion of 21 band-limited streams.*

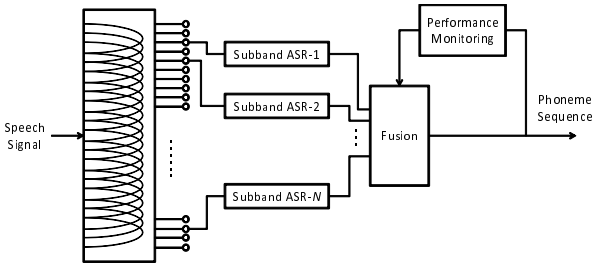

Figure 2: *Schematic diagram of multistream ASR system with feedback control-based fusion: Most reliable (uncorrupted) N streams are selected from all band-limited streams on the basis of performance monitoring and then fused.*

stream speech recognition 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. Several unsupervised techniques were proposed and investigated to select the least corrupted processing streams. Our work is directly based on that carried out by Sharma et al. [6].

The key to the success of multistream ASR lies in performance monitoring that evaluates the performance of individual streams without requiring knowledge about correct answers. Figure 2 depicts a schematic diagram of multistream ASR system with feedback control-based fusion in which multiple reliable streams are selected on the basis of performance monitoring and then fused. The first stage of parallel processing estimates the posterior probabilities of phones 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 feedback from performance monitoring [7, 8, 9, 10]. Along these lines, Mesgarani et al. [7, 8] demonstrated that the best combination of processing streams could be selected on the basis of the autocorrelation of the posterior probability vectors derived from training and testing data.

Variani and Hermansky [9] extended the performance measure to include the Mahalanobis distance between the means of posterior estimates from training and testing data. The experimental results indicated that these criteria for performance monitoring worked reasonably well but required a minimum of four seconds to obtain stable estimates of the probability distribution for posterior data.

3. Subband ASR and Fusion

Our work is based on the two-stage multistream ASR system proposed by Sharma et al. [6] and Li et al. [5] in which the posterior probabilities are derived for each individual parallel subband ASR system and then fused by using merging MLP.

The posterior probabilities of 40 phoneme classes are estimated by three-layered MLP in each band-limited stream in the first processing stage. The features for MLP estimation are derived by frequency-domain linear prediction (FDLP) [11], which characterizes temporal fluctuations in sub-band envelopes. Relatively long temporal spans (500 ms) of FDLP-estimated Hilbert envelopes in each frequency band form the basis of features for the first stage of MLP classification.

We focused on merging MLP to fuse information from individual parallel streams in the present study. We concatenated the logarithm of 40 posterior probabilities obtained from an individual stream and trained an MLP on the concatenated logarithmic posterior features to estimate the final phoneme posterior probabilities [5] that could then be used in a Viterbi search for the best phoneme string [4].

4. Adaptive Selection and Integration of Reliable Processing Streams

The corruption of signals during operation of the system could corrupt some of the processing streams, but depending on the nature of corruption, it could leave some of the streams relatively intact. We attempted to reduce the effect of unreliable (i.e., noise corrupted) streams to improve the robustness of our system against noise by evaluating the reliability of the recognition results in each processing stream by estimating the divergence of each stream output on its training data and on the unknown test.

Speech signals may be corrupted by external sources that were not seen during the training. Therefore, it is not always feasible to cover in the training data as many of these sources as possible [12]. From the aforementioned discussion, the present paper focused on the case of training MLPs in each processing stream on uncorrupted clean speech data, which was the extreme case to evaluate the robustness against unexpected signal distortions. Since the speech messages in the training data and in the test were generally different, we relied on measures of divergence that reflected some general properties of the classifier output computed over relatively long time spans.

The rest of this section describes the GMM-based method of performance monitoring (in Section 4.1), the method of forming multiple processing streams (in Section 4.2), and the method of selecting and integrating reliable processing streams (in Section 4.3).

4.1. Performance monitoring using GMM

Earlier work [7, 9] computed statistics on the root-compressed posterior probabilities of training data and other statistics derived from data during operation. The large divergence between these two statistics indicates unreliable classification. Such methods do not require decisions to be made about classes and

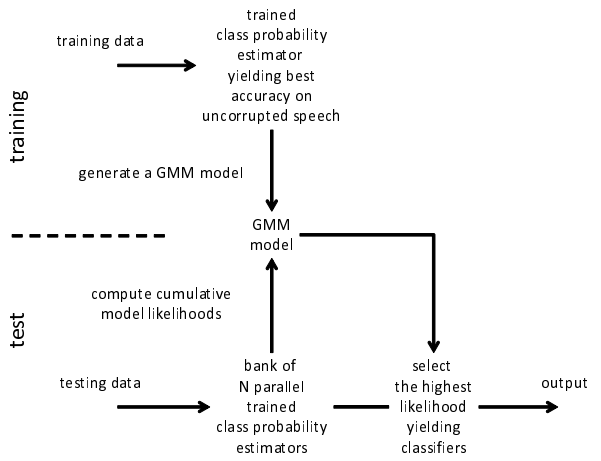

Figure 3: *Schematic diagram of performance monitoring to select reliable processing streams. Class probability estimator applied in training was merging MLP in Fig. 1 yielding best phoneme accuracy on uncorrupted clean speech data.*

they only require knowledge about estimated class probabilities, thus avoiding the need for knowing the ground truth about underlying unknown classes. They have shown that the differences in the autocorrelations and means of posterior probabilities between training and test data are effective for performance measures [7, 9]. However, these measures assume a single Gaussian distribution for the compressed posterior probabilities. It has been reported that reliable estimates of statistics during operation have required at least 4 s of data. This could present problems in faster changing non-stationary noise.

The method we propose here involves building a Gaussian mixtures model (GMM) on the logarithm of output posterior probabilities derived from uncorrupted speech data on which the classifier was trained and evaluating the cumulative model likelihood on some segment of test data. Figure 3 shows the procedure of performance monitoring to select reliable processing streams. A low cumulative likelihood indicates unreliable classifier output. GMM-based evaluation also does not require knowledge of the ground truth, which is similar to methods that employ divergence in statistics. We expect decreasing the time interval necessary to estimate performance by eliminating the assumption of a single Gaussian. It can contribute to dealing with non-stationary noise. The accumulated likelihood for the posterior probabilities obtained from merging MLP in k-th processing stream at frame t is described as:

$$\mathcal{L}(k,t) = \sum_{n=0}^{N-1} \log \sum_{m=1}^{M} w_m^{\mathbf{ref}} \mathcal{N}(\mathbf{p}_{t-n}^k | \mu_m^{\mathbf{ref}}, \mathbf{\Sigma}_m^{\mathbf{ref}}), \qquad (1)$$

where N denotes the number of frames to accumulate likelihood, M is the number of mixture components, $\mathbf{p}_{t-n}^k$ represents the logarithmic posterior probabilities from merging MLP in k-th processing stream at frame $t-n$. Here, $w_m^{\mathbf{ref}}$, $\mu_m^{\mathbf{ref}}$, and $\mathbf{\Sigma}_m^{\mathbf{ref}}$ correspond to the mixture weight, mean, and variance in m-th mixture component of GMM. $\mathcal{N}(\cdot)$ represents a Gaussian distribution.

4.2. Forming processing streams

This subsection briefly describes the two-stage forming of multiple processing streams. Figure 4 outlines the procedure for

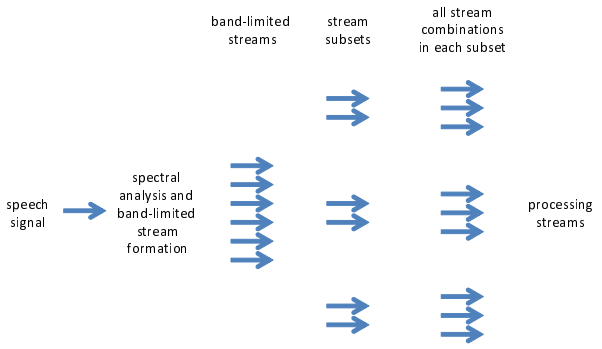

Figure 4: *Procedure for forming processing streams with six band-limited stream system. Six band-limited streams would yield only nine processing streams with procedure described here. Our actual system starts with 21 band-limited streams and yields 381 processing streams.*

forming processing streams. At first, the full frequency is divided into 21 band-limited streams. Then, the processing streams are formed for all non-empty combinations of 21 band-limited streams. However, forming all possible combinations of 21 streams would result in 2,097,151 processing streams. Although this should not present a conceptual problem, it certainly raises some practical computational issues. Therefore, we attempted to reduce the number of processing streams by dividing the 21 band-limited streams into three exclusive subsets of seven streams (7B-1, 7B-2, and 7B-3). In this case, each subgroup resulted in 127 processing streams, yielding 127 × 3 = 381 final processing streams to be evaluated and fused.

4.3. Online stream selection and integration

The procedure for selecting and integrating the reliable processing streams is described as:

1. Select K-best streams from 381 processing streams by thresholding the GMM-based performance evaluation measure applied at the output of each stream as:

$$\mathbf{p}_t^1, \mathbf{p}_t^2, \cdots, \mathbf{p}_t^K = \underset{\mathbf{p}_t^k}{\arg\,\mathrm{nbest}}\, \mathcal{L}(k,t), \qquad (2)$$

   where $\mathbf{p}_t^k$ denotes the logarithmic posterior probabilities obtained from k-th reliable merging MLP at frame t.

2. Fuse K-best posterior probabilities as:

$$\hat{\mathbf{p}}_t = \sum_{k=1}^{K} \alpha_t^k \mathbf{p}_t^k, \quad \sum_{k=1}^{K} \alpha_t^k = 1, \qquad (3)$$

The reliability of k-th processing stream, $\mathcal{L}(k,t)$, is computed at every frame by using previous N frame data, $\mathbf{p}_{t-1}^k, \mathbf{p}_{t-2}^k, \cdots, \mathbf{p}_{t-N}^k$, and reliable processing streams can be adaptively updated at every frame.

We tried the arithmetic mean of posterior estimates, their geometric mean, and majority voting of the classifier decisions for fusion. We ended up using the geometric mean in the present work, i.e., the arithmetic mean of logarithmic posterior features described in Eq. 3 because the preliminary experiments revealed that the geometric and arithmetic mean yielded approximately the same accuracy and they achieved better accuracy than majority voting.

5. Phone Recognition Experiment

5.1. Experimental setup

We used a phoneme recognition system based on the HMM-ANN paradigm [4] that was trained on clean TIMIT data. The car noise and subway noise at SNRs of 5 dB and 15 dB were overlapped with the test data. The speech data were sampled at 16 kHz and quantized into 16 bits. The training data consisted of 3,000 utterances by 375 speakers, a development set (i.e., a cross-validation data set) consisted of 696 utterances by 87 speakers, and a test data set consisted of 1,344 utterances by 168 speakers. The TIMIT database, which was manually labeled using 61 labels, was mapped to the set of 40 phonemes [13]. The development set was mainly used for early stopping in MLP training.

5.1.1. Subband ASR

The FDLP-based acoustic features were converted to 40-dimensional phoneme posterior probabilities by using a three-layered MLP for each of 21 band-limited streams. The acoustic features used were 170-dimensional parameters that were comprised of the 42-dimensional static and 42-dimensional dynamic compression coefficients of FDLPs [11], auditory power, and their frequency derivatives. The number of units in the input, hidden, and output layer corresponded to 170, 1000, and 40.

5.1.2. Multiple processing stream forming

The logarithmic phoneme posterior probabilities from the individual band-limited stream to be fused were concatenated to form input to second stage three-layered merging MLP, which was trained to yield a vector of 40 phoneme posterior probabilities. There were $N_s \times 40$, 1000, and 40 units in the respective input, hidden, and output layers in merging MLP, where N_s denotes the number of streams fused.

5.1.3. Selection and integration of reliable processing streams

Twenty-one band-limited streams were divided into three exclusive sets of seven streams (7B-1, 7B-2, and 7B-3). The seven streams used in the present study were:

- **7B-1:** 3, 5, 7, 9, 12, 16, and 19-th stream
- **7B-2:** 2, 4, 8, 11, 14, 17, and 20-th stream
- **7B-3:** 1, 6, 10, 13, 15, 18, and 21-th stream

The optimal number of mixture components, M used in Eq. 1, was experimentally determined to be three. The number of processing streams selected and integrated, K in Eq. 1, was also experimentally determined to be 70 for uncorrupted clean speech, 40 for speech corrupted by car noise, and 100 for speech corrupted by subway noise. The fusion weights of reliable stream outputs, α_t^k in Eq. 3, were equal for all processing streams fused i.e., $\alpha_t^k = 1/K$ for $k = 1, \cdots, K$.

5.2. Experimental results

We evaluated conventional single-stream ASR systems and the proposed multistream ASR system as:

1. **Single:** Single-stream system using FDLPs as feature parameters. This is equivalent to the system in which the FDLP features are directly taken as inputs at the multi-stream fusion stage in Fig. 1.
2. **21B:** Single-stream system formed by fusing 21 band-limited stream posterior features [5].

Table 1: *Phoneme accuracy (%) of proposed adaptive multi-stream ASR system as function of time intervals necessary for estimating performance with car and subway noise at SNRs of 5 dB and 15 dB. This table includes phoneme accuracy of conventional single-stream ASR systems.*

Method	Interval [ms]	clean	car		subway	
			5 dB	15 dB	5 dB	15 dB
Single	—	66.84	48.63	53.51	39.04	52.78
21B	—	70.08	53.50	59.94	42.38	56.99
PM_GMM	250	70.55	58.03	63.99	42.40	58.37
	500	70.76	58.66	64.49	42.57	59.07
	1,000	70.82	58.84	64.79	42.59	59.11
	2,000	70.99	59.38	65.15	42.82	59.29
	4,000	71.01	59.46	65.42	43.26	59.58
	6,000	71.11	59.69	65.50	43.27	59.55
	8,000	71.09	59.87	65.50	43.08	59.61
	10,000	71.09	59.86	65.48	43.39	59.53

3. **PM_GMM (proposed):** Online estimation and integration of reliable streams using GMM-based performance monitoring.

Table 1 shows the phoneme accuracy of the proposed multistream ASR system based on stream selection and fusion using performance monitoring as a function of the time intervals necessary for estimating performance (250, 500, 1,000, 2,000, 4,000, 6,000, 8,000, and 10,000 ms) with the car noise and subway noise at SNRs of 5 dB and 15 dB. This table also includes the phoneme accuracy of the conventional single-stream ASR system and the single-stream formed by fusing 21 band-limited streams. The single-stream system formed by fusing 21 band-limited streams (**21B**) achieved significantly better phoneme accuracy than that of the conventional full-band ASR system (**Single**).

The proposed system (**PM_GMM**) reduced the phoneme errors in the earlier single-stream system [5] even for the shortest (250 ms) time interval to evaluate performance. In addition, better accuracy was obtained when the time interval to estimate performance was longer. It should be noted that previous work on performance monitoring that used the divergence in statistics [9] under the assumption of the single Gaussian for the logarithmic posterior probabilities raised computational problems due to the data deficiency for short time intervals (e.g., less than 1,000 ms) i.e., it could not stably compute the performance measure such as the autocorrelation matrix of testing data. These results indicate that selection and fusion of reliable streams using GMM-based performance monitoring can efficiently improve robustness against noise.

6. Conclusion

We proposed multistream speech recognition that was adaptive and robust against noise using multiple streams formed by various combinations of band-limited streams and GMM-based performance monitoring. Reliable phoneme posterior probabilities were adaptively selected in the proposed system from the outputs of all possible streams with the GMM-based confidence measure of their performance and then integrated. The proposed system more efficiently reduced errors under noise conditions than the conventional full-band ASR system and the single-stream system without performance monitoring.

7. 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] H. Fletcher, "Speech and hearing in communication," *Krieger, New York*, 1953.

[4] H. Bourlard and N. Morgan, "Connectionist speech recognition: a hybrid approach," *Kluwer Academic Publishers, Boston.*, 1994.

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

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

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

[8] 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.

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

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

[11] 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.

[12] H. Hermansky, "Multistream recognition of speech: dealing with unknown unknowns," *Proc. IEEE*, 2013, *To appear*.

[13] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. M. Doss, "Exploiting contextual information for improved phoneme recognition," in *Proc. ICASSP*, March 2008, pp. 4449–4452.