Adaptive Stream Fusion in Multistream Recognition of Speech

Nima Mesgarani[1,2], Samuel Thomas[2], Hynek Hermansky[2]

[1] Department of Neurological Surgery, University of California San Francisco
[2] The Center for Speech and Language Processing, Johns Hopkins University
nima.mesgarani@ucsf.edu, (samuel,hynek)@jhu.edu

Abstract

A new method to deal with variable distortions of speech during the operation of the system is proposed. First, multiple processing streams are formed by extracting different spectral and temporal modulation components from the speech signal. Information in each stream is used to estimate posterior probabilities of phonemes. Initial values for a weighted integration of these individual estimates are found by normalized cross-correlation of the estimates with the actual phoneme labels on the training data. A statistical model of the final estimated posterior probabilities is used to characterize the system performance. During the operation, the weights in the linear fusion are adapted using particle filtering to optimize the performance. Results on phoneme recognition from noisy speech indicate the effectiveness of the proposed method.

Index Terms: multistream speech recognition, spectrotemporal modulations

1. Introduction

Multistream recognition paradigm for processing of corrupted signals has been studied for more than a decade [1]. In this framework, a number of different representations of the signal are processed and classified independently, for example, different frequency bands [1] or spectrotemporal modulations [2, 3]. Multiple representations allows for a possibility to adaptively alleviate the corrupted channels while preserving the uncorrupted ones for further processing. The paradigm is motivated by results of perceptual experiments carried at Bell Laboratories in the first half of the 20th century [4]. These results showed that the final error of recognition of out-of-context speech sounds is given by the product of errors in the individual sub-bands of the available speech spectrum. This observation is of a significant interest since it may partially explain extreme resistance of human speech communication in presence of variable frequency-localized unexpected disturbances (noises) that are often encountered during decoding of speech messages (Figure 1a).

The notion of heavily parallel multi-stream processing is consistent with the structure of the human hearing system with its relatively small number of neurons at the hearing periphery and at least a couple of orders of magnitude larger number of neurons in higher stages of the hearing system. As evidenced by measured auditory receptive fields of mammalian brains [5], each cortical neuron could in principle represent one individual processing stream with its individual properties. This, together with a hypothesized mechanism for the feedback from the message–decoding cognitive system to acoustic processing, allows for a human listener to adaptively suppress the channels that are heavily corrupted and enhance the relatively clean channels, until the listener believes that the message is being received.

This process is much more difficult to emulate in a machine, which typically does not have means for deciding that "the message is being received". A critical issue is how to dynamically alleviate the corrupted streams in presence of variable disturbances. Some past relevant works used inverse entropy of the classifier output to assess the quality of the stream [6, 7].

In this work, we incorporated dynamic adaptation in our previous system described in [3]. We propose a second-order statistical model of phoneme posterior estimates of each stream and the final output for their characterization. The model derived from the clean data represents the ideal behavior of the system and any decline of performance due to a data corruption is measured by a degree of dissimilarity between clean model and the one derived from the corrupted speech. In this unsupervised method, there is no notion of correct or incorrect classification, all that is measured is the deviation of the output statistics due to the corruption.

In addition, we propose an adaptive fusion technique using Particle Filtering (PF) [8] to adjust the way the posteriograms of streams are fused in order to maximize the similarity of statistics of the fused output to what is expected in the clean condition. Since this filter can be easily implemented in recursive mode [8], it has an added advantage of being capable of processing data in real-time. The initial setup of the fusion module is found by maximizing the correlation of the posterior phoneme estimates from the fusion output with the ground truth provided by phoneme labels in the training data.

2. Stream formation

Ideally, the processing streams should be formed in such a way that a noise affects only some of the streams while the rest of the streams remain relatively unaffected. In this study, we produced streams from the spectrotemporal modulation representation of speech [2, 3]. Starting from the speech signal, we estimated the power spectrum using the magnitude of short-time Fourier transform (STFT) with a typical window of length 25 ms and a frame shift of 10 ms. Critical band energies were then estimated from the power spectrum using bark frequency weights [9].

The spectrotemporal modulation content of the spectrograms was estimated using a bank of 2D Gabor filters [10]. The filters were tuned to different spectral (scale) and temporal (rate) modulations with selectivity to up and down sweeps [10] resulting in a 4D signal that varies along time, frequency, rate and scale (Figure 1b). We observed earlier that in this domain, speech has a distinct spectrotemporal pattern that could often be different from many noises and environmental distortions [11]. Figure 1c illustrates this point using ripple noise that overlaps with speech spectrogram but is mostly confined in only one rate-scale filter (circled in Fig. 1c). Therefore, we formed streams by dividing the spectrotemporal modulations into several subgroups. Out of many possible allocations, here we focused on one such arrangement where upward and downward selective filter were grouped into two sets of low and high scales (figure 1b). In that way, we formed four streams that covered the full range of spectrotemporal modulations.

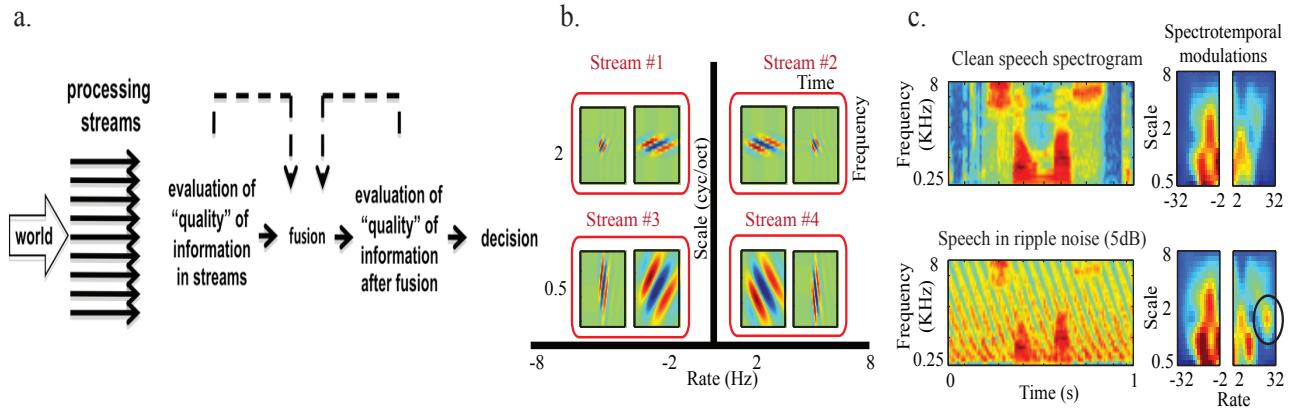

Figure 1. (a) Hypothesized multistream recognition system (b) Spectrotemporal representation of speech estimated by filtering the auditory spectrogram with a bank of 2D Gabor functions tuned to different rate (temporal modulations), scale (spectral modulations) and up-down directions. (c) Spectrogram and rate-scale representation of clean speech and speech in ripple noise. Despite of being overlapping in spectrogram domain, noise is more confined in the modulation domain (circled).

3. Estimation of fusion weights

The fusion of stream posteriograms can be done in several ways [3]. The fusion needs to take into account both the informational relevance of each stream for different phonemes (static) and also be able to change according to the signal-to-noise ratio of each stream (dynamic). For the initial setup of the linear fusion module, we optimize the correlation of estimated posteriograms with the actual labels provided in training data. In this operation, posterior probability of phoneme k at each time frame, $\hat{p}_k$, is estimated using a weighted sum of the posterior probabilities of all phonemes of all streams:

$$\hat{p}_k = \sum_j \sum_i \hat{p}_{i,j} W_{i,j,k} \qquad (1)$$

where $\hat{p}_{i,j}$ is the estimated posterior probability of phoneme i from stream j and $W_{i,j,k}$ is the fusion weight. In order to find the optimal weight, W, we minimize the mean-squared-error between actual and estimated output posteriograms over the training data:

$$e = \sum_t \sum_k \left(p_k(t) - \hat{p}_k(t) \right)^2 \qquad (2)$$

which results in [3]:

$$W = \frac{\hat{P} P^t}{\hat{P} \hat{P}^t} \qquad (3)$$

Intuitively, the above solution is the cross-correlation of estimated posteriograms of each stream with the actual labels, normalized by the autocorrelation of stream posteriograms. The weights W are estimated from the training subset, and then are used to estimate the posteriors of the test data. We showed previously that the linear fusion described above compares well against more complex fusion methods such as MLPs [3]. Next, we show how the

4. Evaluating posteriograms

One way to summarize the behavior of the streams is to measure the statistics of their estimated phoneme posteriogram. We used a second order model (autocorrelation matrix) as defined below:

$$AC_j = \frac{1}{N} \sum_{n=1}^{N} P_j(n) P_j(n)^T \qquad (4)$$

where $P_j(n)$ is the vector of posterior estimates of stream j at frame n (size = 39) and AC_j is the autocorrelation matrix for stream j (size = 39 by 39) (Figure 2). The diagonal elements of this autocorrelation matrix reflect the estimated occurrence frequency of each phoneme and the off-diagonal values correspond to the co-activation of different phoneme posteriors.

Autocorrelation does not tell anything about whether the posterior estimates were correct, it merely reflects the first (diagonal) and second order (off-diagonal) statistics of the estimated posteriograms. However, the off diagonal elements reflect confusions among phoneme classes because an ideal posteriogram has only one phoneme active at each time instance. This is shown in Figure 2a top, which shows two clean posteriograms estimated from streams 1 and 4 along with the corresponding autocorrelation matrices. As circled in Figure 2a, posterior estimates of vowel and consonant groups indicate more within-class correlation indicating their higher confusability.

4.1. Evaluating distortions of the posterior estimates

The autocorrelation matrix computed from posteriograms of undistorted speech summarizes the performance of each stream in clean condition. Any distortion of the posteriogram due to any factor results in the change of this statistics. Thus, computing a measure of similarity between the autocorrelation matrices derived from the clean and the corrupted signals (regardless of the nature of the corruption) indicates the degradation of the stream due to the corruption. We used the Pearson's correlation [12] to quantify this similarity, defined as:

$$r = \frac{AC_{clean} AC_{noisy}}{\left\| AC_{clean} \right\| \left\| AC_{noisy} \right\|} \qquad (5)$$

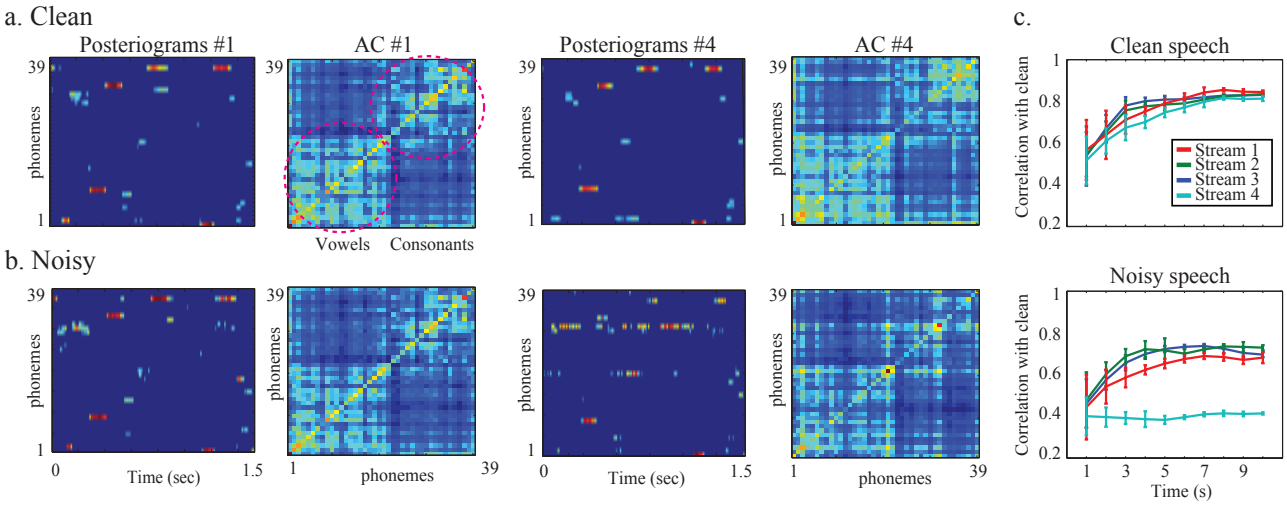

Figure 2. Example posteriograms and their autocorrelation matrices for clean (a) and noisy speech (b), estimated from streams 1 (left) and 4 (right). Stream 4 degrades more severely in ripple noise (b, right). Red circles highlight the within-class confusability of vowels and consonants. (c) Effect of posteriogram duration on stability of correlation measure. Correlation coefficients are stable after about 4 seconds.

where AC_{clean} denotes clean autocorrelation values (clean statistics, vectorized) and AC_{noisy} is the autocorrelation after the corruption. This measure proved to be an effective predictor of streams recognition accuracy [13].

An important practical issue is the time needed to obtain a reliable estimate of AC and therefore, r (equation 5). This is important in realistic test situations where systems need to adapt quickly to changes in the environment. For the clean model, AC_{clean}, we used all the data available for the training. However, during the operation, a reasonable estimate of AC_{noisy} may need to be computed only from a limited amount of data. Figure 2c shows the correlation for the four streams in clean (top) and ripple additive noise conditions (bottom) as a function of time span used for this estimation (N in equation 4). Error bars in Figure 2c were estimated from 25 independent segments of posteriograms. As Figure 2c shows, the correlation with clean stabilizes after approximately 4 seconds making it suitable in most practical situations where the noise statistics change gradually.

5. Adaptive fusion

Perhaps the most crucial part of a robust multistream speech recognizer is the ability of the system to adaptively adjust the way streams are fused in order to reduce the effect of nonstationary corruptions. Here we investigated adaptation of the fusion weights (W) using a Particle Filter [8] such that to maintain the statistical similarity of the overall output posteriogram to clean (r from equation 5).

Particle Filter requires two probabilistic models: state model and observation model. The state model describes the evolution of the system from its past state whereas the observation model relates the observations to the current state of the system. In our implementation, the fusion weights (W) constitute the state of the system, and the correlation measure described in equation 5 forms our observation. State transition is modeled as:

$$W_t = W_{t-1} + V_t \qquad (6)$$

where V_t is the system noise and assumed to be *iid* noise process. The state of the system is therefore Markov with transition probability distribution assumed to be normal with equal mean and variance ($N\left(W_{i,j,k}, W_{i,j,k}\right)$). This choice of variance ensures that the PF concentrates on the weights that were significant for estimation of posterior probability of each phoneme in clean. The basic operation of PF in each iteration consists of producing a cloud of M particles by sampling M times from the current probability distribution of state, $N\left(W_{i,j,k}, W_{i,j,k}\right)$. Each particle (new weight) is then used in the fusion to estimate the output posteriogram (equation 1). Using equations 4 and 5, we rank each of the M particles based on the goodness of their corresponding posteriograms (r from equation 5). The final step is updating the state of the system (W) by taking the average of the M particles weighted by their ranks. In effect, this procedure changes the fusion weight of each phoneme of each stream ($W_{i,j,k}$) such that the statistics of the output estimated posteriogram (AC_{noisy}) becomes more similar to the clean (AC_{clean}). This objective is achieved partly by suppressing the weights given to noisy estimates as shown in figure 3. Top and bottom rows in figure 3 correspond to two ripple noises with spectrotemporal characteristics that corrupt some streams more than others. Black bars show the decreased recognition accuracy of four classes of vowels, nasals, plosives and fricatives of each of the four streams. As figure 3 shows, ripple noise 1 (top row) degrades plosive recognition in streams 1,2 significantly more than in streams 3,4. Ripple noise 2 on the other hand (bottom row) degrades plosive and nasal recognitions in streams 3,4 while the fricatives are more degraded in streams 1,2. Red bars in figure 3 show the percentage of change in the fusion weights for each phoneme class after adaptation. The weights change in order to inhibit the contribution of more noisy channels while enhancing the less noisy ones, therefore improving the overall recognition accuracy of the system.

6. Experimental results

We conducted speaker independent phoneme recognition experiments to test the effectiveness of the proposed methods. We used a phoneme recognition system based on the Hidden Markov Model – Artificial Neural Network (HMM-ANN) paradigm [14] trained on clean speech using TIMIT database. The training data set consists of 3000 utterances from 375 speakers, cross validation data set consists of 696 utterances from 87 speakers and the test set consists of 1344 utterances from 168 speakers, sampled at 16 KHz. We estimated the

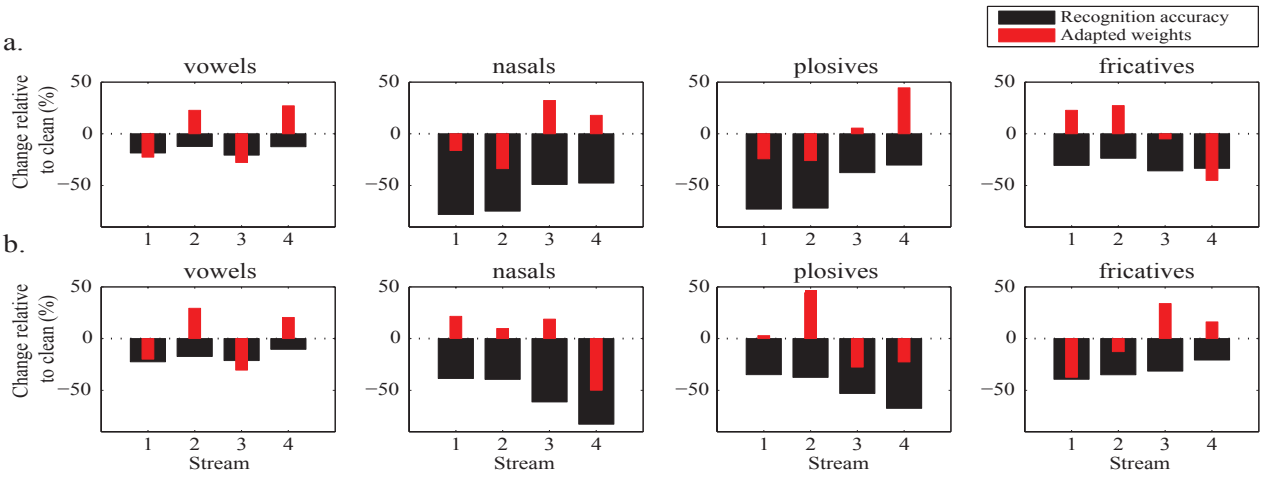

Figure 3. Percentage of change in recognition accuracy (black bars) and the fusion weights after adaptation (red bars) for (a) ripple noise 1 and (b) ripple noise 2. The streams that are more degraded for each phoneme class have decreased weight, while the less noisy streams have enhanced contribution to the output phoneme posteriograms estimate.

posterior probability of phonemes (set of 39 [15]) for each stream using a single hidden layer Artificial Neural Network. The final posteriogram of the system was estimated by combining the posteriograms of all streams using equation 1. Table 1, top row shows the pre-adaptation phoneme recognition accuracy in clean and 7 additive noises from Noisex database at 20dB SNR. Bottom row shows the phoneme recognition accuracy in noise after the fusion weights were adapted to noise using the particle filter. On average, we observed 13.8 % relative error reduction in noise after adaptation.

	Clean	Ripple 1	Ripple 2	Babble	Jet	F16	M109	White
Before adaptation	**62.6**	**37.1**	**33.6**	**42.1**	**32.4**	**35.5**	**41.6**	**35.9**
After adaptation	**62.6**	**39.9**	**39.7**	**46.4**	**38.0**	**41.1**	**47.2**	**40.3**

Table 1. Phoneme recognition accuracy in clean and noise.

7. Discussion

A successful multistream speech recognition system requires three basic elements:
(1) Formation of multiple streams of information that are selective enough to avoid corruption of all streams in noise, and convey enough information for a successful decoding the input from only a subsets of them (2) A way to assess the quality of each stream and the overall system in different conditions (3) An adaptive fusion that combines the streams in a way that minimizes the effect of noisy channels.

In this paper, we presented one possible solution to each of these problems: (1) We propose to use spectrotemporal processing streams based on our knowledge of mammalian auditory cortical processing [5] (2) A method where the similarity of second order statistics of posteriogram with clean condition is used to evaluate the quality of the system output during the operation on possibly corrupted speech (3) This similarity measure can then be used by a Particle Filter to dynamically adjust the streams fusion, emulating the hypothesized process in human decoding of noisy signals, briefly sketched in the Introduction.

We have shown in the past that increasing the number of streams results in better recognition accuracy in clean ([3]). The future work includes extending the adaptive framework to such cases where more constraints on the transition of system may be necessary to achieve a stable adaptation.

8. Acknowledgment

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015 and the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of DARPA, IARPA or its Contracting Agent, the U.S. Department of the Interior.

9. References

[1] Hermansky, H., S. Timberwala, and M. Pavel. *Towards ASR on partially corrupted speech*. in *ICSLP*. 1996: IEEE.
[2] Zhao, S.Y., S. Ravuri, and N. Morgan. *Multi-stream to many-stream: Using spectro-temporal features for asr*. in *Interspeech*. 2009. Brighton, UK.
[3] Mesgarani, N., S. Thomas, and H. Hermansky, *A Multistream Multiresolution Framework for Phoneme Recognition.* Interspeech, 2010. Makuhari, Japan.
[4] Fletcher, H., *Speech and hearing in communication.* New York: Van Nostrand, 1953.
[5] Mesgarani, N., et al., *Phoneme representation and classification in primary auditory cortex.* J Acoust Soc Am, 2008. 123(2): p. 899.
[6] Okawa, S., E. Bocchieri, and A. Potamianos. *Multi-band speech recognition in noisy environments*. in *ICASSP*. 1998: IEEE.
[7] Valente, F. and H. Hermansky. *Combination of acoustic classifiers based on dempster-shafer theory of evidence*. 2007: IEEE.
[8] Arulampalam, M.S., et al., *A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking.* Signal Processing, IEEE Transactions on, 2002. 50(2): p. 174-188.
[9] Hermansky, H., *Perceptual linear predictive (PLP) analysis of speech.* J. Acoust Soc Am, 1990. 87(4): p. 1738-1752.
[10] Chi, T., et al., *Multiresolution spectrotemporal analysis of complex sounds.* J Acoust Soc Am, 2005. **118**(2): p. 887-906.
[11] Mesgarani, N., et al., *Denoising in the domain of spectrotemporal modulations.* EURASIP J Aud Sp Proc. 2007. (3): p. 3.
[12] Rodgers, J. and W. Nicewander, *Thirteen ways to look at the correlation coefficient.* Am Statistician, 1988. **42**(1): p. 59-66.
[13] Mesgarani, N., S. Thomas, and H. Hermansky, *Toward optimizing stream fusion in multistream recognition of speech.* J Acoust Soc Am Express Letters, in press.
[14] Bourlard, H. and N. Morgan, *Connectionist speech recognition: a hybrid approach*. 1994: Springer.
[15] Halberstadt, A. and J. Glass. *Heterogeneous Acoustic Measurements for Phonetic Classification*. 1997: ISCA Eurospeech.