PERFORMANCE MONITORING FOR ROBUSTNESS IN AUTOMATIC RECOGNITION OF SPEECH

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

[1] The Johns Hopkins University, Baltimore, MD
[2] University of California, San Francisco

ABSTRACT

A new method to deal with an unexpected harmful variability (noise) in speech during the operation of the system is reviewed. The fundamental idea is to derive in the training phase statistics of the system output for the data on which the system was trained and adaptively modify the system so that statistics derived during the operation are similar. 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 speech sounds (posteriogram) in each stream, and these estimates are fused to derive the final posteriogram. The autocorrelation matrix of a modified final posteriogram is adopted as the measure that summarizes the system performance. Initial setup of the fusion module is found by cross-correlating the probability estimates with phoneme labels on training data. During an operation, the matrix derived on the training data serves as the desirable target and the fusion module is modified to optimize the system performance. Results on phoneme recognition from noisy speech indicate the effectiveness of the method.

1. THE PROBLEM

One of critical differences between information processing by biological systems and by machines is in rapid decline of performance of a machine, exposed to unexpected harmful signal variability (noise).

A biological system must make sense of relevant information in the environment that surrounds it. The environment is typically cluttered by information from other irrelevant sources of a harmful variability. The organism must be able to focus on the relevant information. Understanding how is this done and emulating such ability in a machine would greatly increase the machine utility in most practical information extraction tasks.

2. CURRENT STATE-OF-THE-ART IN MACHINE LEARNING

Machine learning approaches build a model of the world by optimizing performance on some training data. This assumes that the world will not change in the future. However, this cannot be guaranteed. As many would agree, our world is full of "unknown unknowns" [9], some of them possibly harmful, and among them are also signal distortions not seen in the training data.

One approach to addressing the unexpected noise is to derive features that are invariant to noise. This seems promising but is difficult for many noise types. Another approach is to adapt the model so that it better fits the new incoming data. The adaptation needs to happen reasonably fast, thus acting on relatively limited amounts of new data. Supervised adaptation requires labeled new data, existing unsupervised approaches assume that increasing model likelihood on the limited adaptation data ensures reasonable recognition performance in new situation. This may not apply for truly unexpected noise types.

In spite of significant efforts in those two directions, it appears that some more fundamental flaw in machine design needs to be corrected to succeed in dealing with unexpected sources of the unwanted variability.

3. OUR MOTIVATION

Human auditory cortex contains several millions of cortical neurons, each neuron in principle providing for a separate stream in processing an incoming auditory stimulus. It is very likely that reasonable corruptions of the signal (reasonable in the sense that some of the information can still be decoded by human listener) affect various cortical processing streams differently.

As evidenced in results of physiological (e.g. [1]) and psychophysical (e.g. [2]) experiments, cognition may be able to provide for a metacognitive feedback loop from the output of the system. For example, recent studies suggest the anterior cingulated cortex may play a role in generating the error related signal and communicate it to the rest of the cognitive system. While [1] shows that components in recorded EEG signals may indicate the certainty of the decision made by subjects, that correlates with the stimulus degradation, [2] shows the animal can asses the certainty of their decision and use it to maximizes the expected reward. This metacognitive performance monitoring process may be able to adaptively suppress the streams that are heavily corrupted and enhance the relatively clean streams, until the listener believes that the message is being received.

Providing similar mechanism in engineering schemes for machine information extraction could be a starting point for a new generation of machine learning techniques that could alleviate the excessive fragility of machine performance in presence of unexpected signal corruptions.

4. PERFORMANCE MONITORING IN MACHINE LEARNING

First, we postulate that the classifier performance is optimal for the data on which the classifier was trained, and any deviation from this ideal condition only degrades the performance.

Further, we propose that the classifier performance can be characterized by computing statistics at the output of the classifier. Such statistics do not require making decisions about the classes and only require the knowledge about estimated class probabilities, thus avoiding the need for knowing the ground truth about the underlying unknown classes. This is consistent with reality in deriving information from sensory inputs: we do not know what the truth is, all we know are its estimates.

Thus, comparing statistics derived from the classifier output on the "good" training data and the "corrupted" data that are available in the operation, may give some indication how corrupted the real data are.

5. PERFORMANCE MONITORING FOR ADAPTIVE FUSION IN MULTICHANEL MACHINE RECOGNITION

In the multistream classifier discussed here, the estimates of corruption in the individual streams can be used to decide which streams to emphasize or de-emphasize during the fusion (nonrecursive monitoring). Alternatively, statistics derived at the output of the fusion can be used in global optimization of the system by attempting to make the statistics derived from the "corrupted" real data and the "good" training data more similar (recursive monitoring).

5.1. Estimating Output Statistics

Mesgarani and his colleagues [3] suggest to compute second order statistics at the output of a neural net based classifier that estimates from short segments of speech signal a sequence of equally-spaced vectors of posterior probabilities of underlying phonetic classes (posteriogram).

They estimate the posterior probability of phonemes (set of 39 [4]) for each stream using a single hidden layer multi-layer perceptron artificial neural network (ANN). The accuracy of phoneme posterior probability estimation in each individual stream is less than that of a baseline system, since each stream carries only a part of the available information. One way to measure statistics of the posteriogram regardless of its duration, is computing an autocorrelation matrix

$$AC_n = \frac{1}{T}\sum_{t=0}^{T} F[\mathbf{R_n(t)}]F[\mathbf{R_n(t)}]^T \qquad (1)$$

where $\mathbf{R_n(t)}$ is the vector of posterior estimates of stream n at time t, and F[.] is some monotonic compressive function (currently the fifth root). The diagonal elements of this autocorrelation matrix reflect the occurrence frequency of each phoneme and the off-diagonal values corresponding to the co-activation of *different* phoneme posteriors. It 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.

5.2. Estimating Distortions

The autocorrelation matrix computed from the undistorted signal indicates the behavior of the stream on such ideal data. Any additional distortion of the signal results in the change of the statistics that this matrix describes. Thus, computing a measure of similarity between the autocorrelation matrixes derived from the clean signal and from the signal corrupted by any means indicates the degradation of the stream due to the distortion. This measure is

$$r = \frac{AC_{clean}AC_{corrupted}}{\left\|AC_{clean}\right\|\left\|AC_{corrupted}\right\|} \qquad (2)$$

where AC_{clean} denotes uncorrupted autocorrelation matrix (clean model) and $AC_{corrupted}$ is the matrix after the corruption.

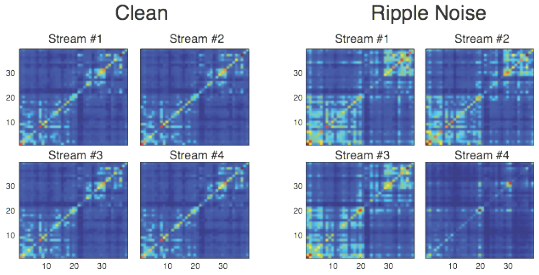
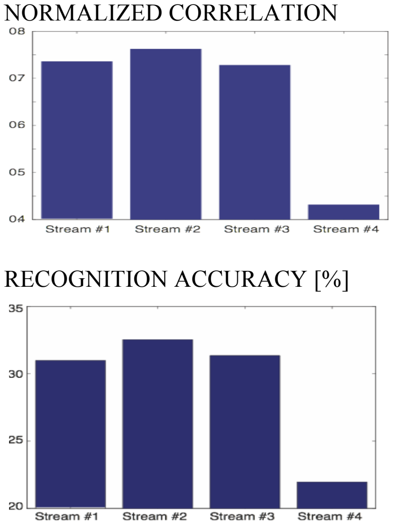


Fig. 1 Autocorrelation matrixes estimated for the clean (left part of the figure) and distorted (the right part) signals in four different processing streams. The distortion was mainly affecting the stream #4

The clean estimate, AC_{clean} used all the data available for the training. However, during the operation, the $AC_{corrupted}$ may need to be estimated only from a limited amount of data. Thus, an important practical issue is the duration of posteriogram needed to obtain a reliable estimate of $AC_{corrupted}$. To test the dependence of similarity measure on the duration of posteriogram (T in equation 1), [3] calculated $AC_{corrupted}$ of the four streams in ripple additive noise while increasing T. The correlation with the matrix clean seemed to stabilize after ~4 seconds making it suitable in most practical situations, where the noise statistics changes are relatively gradual.

As illustrated in Fig. 3 (adopted from [3]), the cross correlation predicts well the recognition accuracies in the individual streams. (In this example, the most corrupted stream was the stream number 4).

Notice that for evaluating the quality of information in the individual processing streams, there is no need of knowing what the result of the classification is, *all that is required to know is how the system works on the training data.*


NORMALIZED CORRELATION

RECOGNITION ACCURACY [%]

Fig. 3 Pearson's correlation (Eq. (2)) between correlation matrices derived from clean and noisy posterior estimates in four different processing streams and the phoneme recognition accuracies in these four streams (adopted from [3]).

For the labeled training data, the correlations between matrices derived on class labels and the posterior probability estimates from the training data can be used for estimating relative efficiencies of the individual processing streams that can be used for the initializing of adaptive linear fusion [7].

5.3. Monitoring Individual Streams

One way towards increasing robustness in the multistream system is to indicate corrupted streams prior to the fusion and to use this knowledge in the fusion. Earlier works applied for this purpose several techniques for estimating the classifier confidence such as difference among several top posterior estimates or noise estimates in the streams [5], or entropies of the classifier output [6].

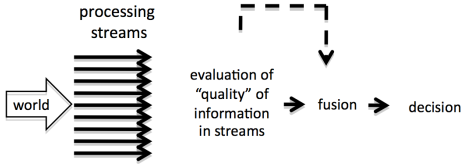


Fig. 4 Nonrecursive fusion control

Recently, picking up the best performing streams based on the criterion (2) has been also applied for eliminating the corrupted streams in linear fusion [3].

5.4. Monitoring The Recognizer Output

In a more global framework, one can monitor the performance of the overall output of the system instead of the individual processing streams. This eliminates the need for performance monitoring on each processing stream by focusing on the contribution of streams to the final output. One advantage of such framework is its ability to optimize the fusion by taking into account the changes in statistics of all streams at once, which can be more effective than examining streams independently.

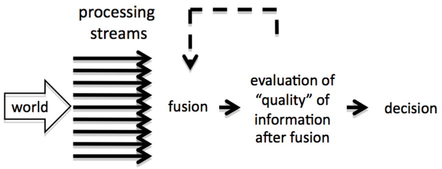


Fig. 5 Recursive fusion control

The proposed feedback loop only requires evaluating the final output from the information processing path and a single stream communication link used in modifying the fusion module without the necessity of communication of information between the fusion and all processing streams.

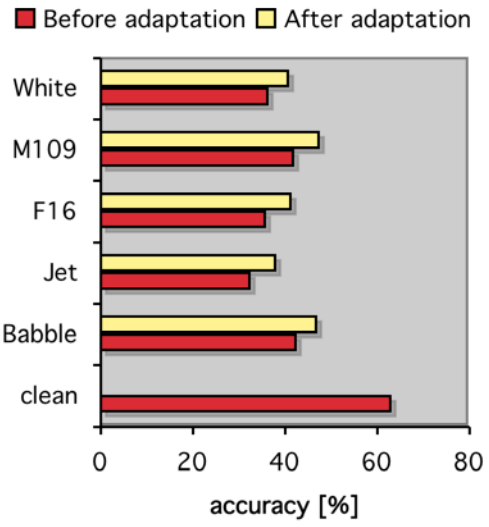
Mesgarani et al [3] demonstrated that the technique can be used for choosing the best combination of processing streams in neural net based fusion. However, it requires training of the fusion on each possible $n!/k!(n-k)!$, $\forall k=1\ldots n$ combination of streams, which presents obvious engineering difficulties for larger number of processing streams.

More recently Mesgarani et al [7] came with more practical linear fusion based technique where the fusion is adaptively modified using particle filtering techniques. Appropriate initial weights for a weighted summation of individual estimates are found by cross-correlating the estimates with the phoneme labels on the cross-validation (or training) data. The autocorrelation matrix *of the final weighted probability estimates* is adopted as the measure that summarizes the system performance during the operation. The weights in the linear fusion are adapted using particle filtering to optimize the system performance. In effect, the procedure changes the fusion weight of each phoneme in each stream so that the statistics of the output, computed using Eq. (1) are more similar to the statistics derived on the clean data.

4. RECOGNITION EXPERIMENT

As reported in [7], this technique was tested on phoneme recognition of TIMIT sentences (3000 utterances from 375 speakers in the training, 696 utterances from different 87 speakers as the cross-validation data, and 1344 utterances from 168 speakers in the test, all sampled at 16 kHz). The system was always trained on the original clean TIMIT data, the test utterances were corrupted by adding various noises from the Noisex database at levels that noticeably degraded the recognizer performance. As shown in the Table I, the procedure described in the paper [7] resulted in noticeable improvement of performance in all noise conditions, with the average relative improvement of 13.8 %.

TABLE I. Phoneme recognition accuracies in recognition of noisy speech (noise types indicated next to the bars). Adopted from [7].



5. CONCLUSIONS

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 performance of the system in different conditions (3) An adaptive fusion that combines the streams in a way that minimizes the effect of noisy streams so that the performance of the system improves.

In this paper, we provided motivation for the multistream approach and summarized some solutions (described in more detail elsewhere [3],[7]) to each of the problems mentioned above:

(1) The use of spectrotemporal processing streams based on our knowledge of mammalian auditory cortical processing [8] (2) The similarity of second order statistics of posteriograms from training and operational conditions is used to evaluate the quality of the system output during the operation (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.

The proposed solutions have been shown to be effective in phoneme recognition of speech that was artificially corrupted by several real-world noises.

ACKNOWLEDGMENT

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015. 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 or its Contracting Agent, the U.S. Department of the Interior.

REFERENCES

[1] Sheffers and Coles, Performance Monitoring in Confusing Word, J. Exp. Psychology 26 (1)

[2] Smiths and Washburn, Uncertainty Monitoring and Metacognition by Animals, Current Directions In Psychological Science, Vol 14, No 1, 2005

[3] N. Mesgarani, S. Thomas, and H. Hermansky, Towards optimizing stream fusion, J. Acoust. Soc. Am. Volume 130, Issue 1, pp. EL14-EL18 (2011)

[4] A. Halberstadt and J. Glass, Heterogeneous Acoustic Measurements for Phonetic Classification 1, 1997

[5] H. Hermansky and S. Tibrewala and M. Pavel, Towards ASR on partially corrupted speech, in Proc. of ICSLP'96, pp. 462-465, 1996

[6] S. Okawa, E. Bocchieri, and A. Potamianos, Multi-Band Speech Recognition in Noisy Environments, in Proc. of ICASSP-98 , Seattle, Washington, USA, May 1998

[7] N. Mesgarani, S. Thomas, and H. Hermansky, Adaptive Stream Fusion in Multistream Recognition of Speech, Proc. Interspeech 2011

[8] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, Phoneme representation and classification in primary auditory cortex, Acoust. Soc. Am. Volume 123, pp. 899-909, 2008

[9] R. Furlong, Clausewitz and Modern War Gaming: losing can be better than winning, Air University Review, July-August 1984