MEAN TEMPORAL DISTANCE: PREDICTING ASR ERROR FROM TEMPORAL PROPERTIES OF SPEECH SIGNAL

Hynek Hermansky [†‡]*,Ehsan Variani*[†] *and Vijayaditya Peddinti*[†]

[†]Center for Language and Speech Processing,
[‡]Human Language Technology Center of Excellence,
Johns Hopkins University, USA
{hynek,variani,vijay.p}@jhu.edu

ABSTRACT

Extending previous work on prediction of phoneme recognition error from unlabeled data that were corrupted by unpredictable factors, the current work investigates a simple but effective method of estimating ASR performance by computing a function $M(\Delta t)$, which represents the mean distance between speech feature vectors evaluated over certain finite time interval, determined as a function of temporal distance Δt between the vectors. It is shown that $M(\Delta t)$ is a function of signal-to-noise ratio of speech signal. Comparing $M(\Delta t)$ curves, derived on data used for training of the classifier, and on test utterances, allows for predicting error on the test data. Another interesting observation is that $M(\Delta t)$ remains approximately constant, as temporal separation Δt exceeds certain critical interval (about 200 ms), indicating the extent of coarticulation in speech sounds.

Index Terms— error-rate prediction on unknown data, phoneme classification, automatic recognition of speech

1. INTRODUCTION

In many practical applications, it would be very useful to be able to predict the classifier performance on unknown test data even when the answers are not known *a priori*. We propose that this may be possible and are encouraged by the fact that human listeners and some higher-level animals demonstrate this ability ([1][2]). Some of our prior work [3][4] approached this problem by comparing statistics of instantaneous estimates of posterior probabilities of speech sounds (estimated by trained artificial neural net (ANN)) derived on training data and in the test. In this work we propose application of one temporal-domain technique, that we previously used for estimating length of information-bearing element in speech [5], for predicting error rates of phoneme recognizer in unseen noisy environments.

The next section reviews related previous work. Subsequent section describes the proposed technique. Application of the technique using two types of features, the conventional PLP cepstral features, and the data-derived posterior features, is discussed next. The last section discusses the results and concludes our findings.

The research presented in this paper was partially funded by the DARPA RATS program under D10PC20015, and the JHU Human Language Technology Center of Excellence. 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 the JHU HLTCOE.

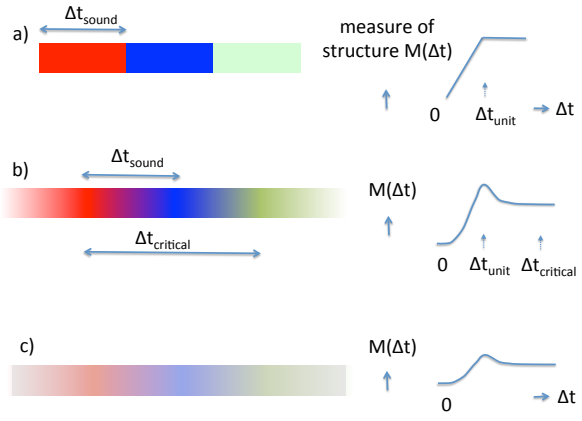


Fig. 1: Cartoon illustration of coding of information in speech and its effect on the mean temporal distance $M(\Delta t)$ of Eq.1. a) Idealized situation b) Coarticulated units c) Effect of stationary corruptions of the signal

2. RELATED PREVIOUS WORKS

2.1. Predicting error in recognition

The fundamental premise is that the *ASR system can never work better than it does on the data on which it was trained.* In this ideal situation, the data that the system encounters during its operation come from the same distribution as do the data on which the system was trained and the system is at its best.

Earlier works in predicting test error ([3][6][7][4]) proposed comparison of a second order statistic $AC = 1/T \sum_{i=1}^{T} P(j)P(j)'$ where $P(j)$ is a vector of 10th root-compressed phoneme posteriors, and T indicates the time span over which the statistic is evaluated. The basic premise is that a classifier is at its best when applied to its training data. Deviations from stochastic regularities derived from the training data degrade its performance. Therefore, large divergence between these two statistics indicate possible degradation of the classifier performance. Divergence, between statistics derived on training data and segments of the test data, $M_{SD} = divergence(AC_{train}, AC_{test})$ then indicates how far the test data deviate from the train data. Several measures of di-

vergence were investigated and were shown to correlate with the observed recognition accuracies [7][4]. Subsequently, they were applied in adapting multi-stream phoneme recognizer to previously unobserved noise in the test data.

The current technique utilizes **temporal properties** of speech, as consisting of sequences of information-bearing speech sounds, expressed in speech features. This differentiates it from earlier approaches, which used statistics of **instantaneous** classifier output.

3. PROPOSED PERFORMANCE MONITORING TECHNIQUE

abelsec:proposed

3.1. Coding message in speech as reflected in $M(\Delta t)$

Speech messages are coded in sequences of speech sounds [1] . A reasonable assumption is that feature vectors describing the speech signal should be similar within each sound and different across sounds. Some time ago [5] we were interested in deriving a typical extent of sound coarticulation and proposed a measure that evaluated a mean temporal distance of features over some interval as a function of time-span Δt between two feature vectors in running speech,

$$M(\Delta t) = \frac{\sum_{i=1}^{T-\Delta t} D(P_t, P_{t+\Delta t})}{T - \Delta t} \quad (1)$$

where D is distance between two feature vectors P_t and $P_{t+\Delta t}$.

Notice that there is no need for labelled data and no need for knowing what the sounds are. The technique is applied directly on any non labelled and/or non transcribed data.

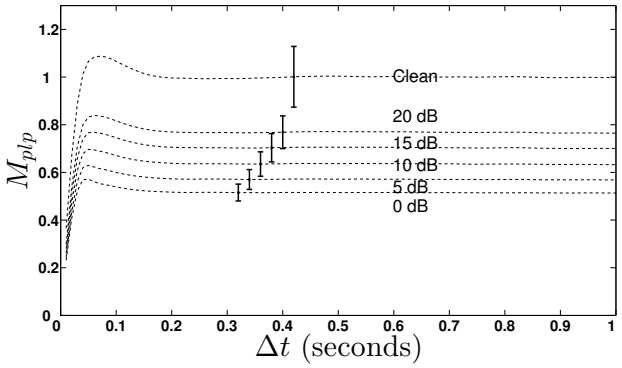
It turns out that, besides evaluating average extent of the sounds, this measure could be also used to evaluate how similar or different (in average) these sounds are. For high quality speech, the sounds are sufficiently different. When the speech gets corrupted by stationary or slowly varying distortions, these distortions start dominating the signal and the sounds become more similar.

The situation is illustrated in Fig.1. The upper part of the figure illustrates an idealized situation where feature vectors in each sound are stationary and all sounds are of equal lengths. The $M(\Delta t)$ increases linearly up to Δ_{sound}, which indicates the length of the sound, and then it stays constant at the value of average divergence between speech sounds. In reality, illustrated in the middle part of Fig1, the feature vectors are not stationary but gradually change due to coarticulation in speech production, and the sounds might be of different lengths. Still, the $M(\Delta t)$ increases gradually with Δt up to the $\Delta t_{critical}$, which indicates the longest time span for which the two feature vectors are guaranteed to be coming from different sounds, i.e. the longest extent of sound coarticulation. $M(\Delta t)$ may exhibit a peak at Δt_{unit}, which is the average time span between unit centers, when features in different units are in average most dissimilar to each other.

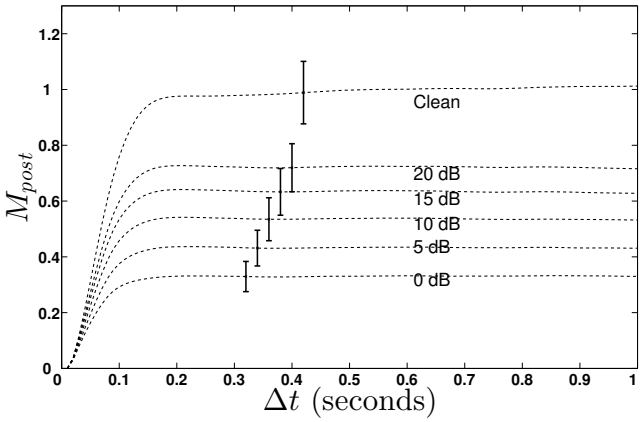
3.2. The effect of signal distortions

The particular shape of $M(\Delta t)$ is dependent on many factors that characterize the data. Most relevant for our current application is that stationary distortions of the signal make all speech units more similar to each other, impairing information-bearing capacity of the speech sounds. Diminished differences in speech sounds show in decreased values of $M(\Delta t)$.

[1] We intentionally avoid here the term "phoneme" or "phone". Our measure merely indicates the presence or absence of structure in the signal that could bear an information. An assumption is that there is are a number of different similar-length interleaved information-carrying sounds. Our measure evaluates their difference in a given feature space and estimates their extent.



(a) $M_{plp}(\Delta t)$: PLP cepstra computed using Euclidean distance



(b) $M_{post}(\Delta t)$: Phone Posterior Vectors computed symmetric KL divergence

Fig. 2: $M(\Delta t)$ curves for in Babble noise at 5 different SNRs, along with variance, using two different MTD formulations

3.3. The effect of different feature representations

With an appropriate divergence measure, $M(\Delta t)$ can be computed using any feature representation of speech. The upper part of Fig.2 shows means and variances of $M(\Delta t)$ curves computed from individual TIMIT sentences using PLP features with Euclidean distance. The lower part of the Fig.2 shows the similar curves but $M(\Delta t)$ computed using phoneme posterior features with symmetric Kullback-Leibler distance.

To facilitate the comparison, the curves were normalized with respect to their values at $\Delta t = 200ms$ in clean conditions. As apparent, PLP-derived curves exhibit peaks at around 70 ms, which corresponds to average spacing between center of neighboring phonemes in the TIMIT data, and they flatten at approximately constant value after about $\Delta t > 200ms$, indicating the main effect of phoneme coarticulation in the TIMIT data. The posterior-derived curves do not exhibit the peak at the average phoneme center spacing. This can be understood since the artificial neural net classifier applied in this experiment was trained to deliver similar features through the

whole length of a phoneme.

3.4. Predicting classifier performance

As implied above, evaluating appropriate elements of $M(\Delta t)$ allows for estimating level of speech distortion, and comparing the $M(\Delta t)$ curves for different data may indicate differences in distortions in the data. In particular, the maxima of $M(\Delta t)$ are proportional to the SNR of the signal, with an offset depending on the type of noise. Following the discussion in Sec.2.1, differences in $M(\Delta t)$ derived from the training and the test data could predict recognition accuracy.

Evaluation of appropriate segment of $M(\Delta t)$ curves yields the measure μ, which indicates how well the data is structured into different segments, which for speech could be speech sounds, related to phonemes of language.

In our case here, the sum of the curve in the region $\Delta t \in < 200ms, 800ms >$ *i.e.,*

$$\mu = \sum_{\Delta t=200ms}^{800ms} M(\Delta t)., \tag{2}$$

indicates level of noise in the given speech data. Further, the relative difference between μ_{test}, derived on a segment of the test data, and μ_{train}, derived on the entire training data, $(\mu_{train} - \mu_{test})/\mu_{train}$ predicts the degradation in performance due to noise.

4. EXPERIMENTAL RESULTS

Experiments done using a phoneme recognition system, trained on the TIMIT database, indicate performance of measure, M, in predicting error increase due to noise. Details of the experimental setup are available in [4]. Recognition error of this phoneme recognition system was measured on the test set of TIMIT database, consisting of 1344 utterances from 168 speakers, corrupted with a variety of noises from NOISEX-92 database [8]. The interval T used in $M(\Delta t)$ computation was, length of the entire training set for $M_{train}(\Delta t)$ and length of each TIMIT sentence, for $M_{test}(\Delta t)$.

4.1. Correlations between predicted performance and relative increase in error of the recognition

As the noise level increases, the error also increases. Relative increases in error plotted against the relative changes in distortion measure, for two different noise conditions (babble noise and buccaneer(I) noise), at varying signal-to-noise ratios, are shown in Fig. 3,4 and 5. Red, blue and green colors are used to represent data corresponding to babble noise, buccaneer(I) noise and the reference clean condition.

The ellipses indicate the contour (at level 0.2) of the Gaussian fit to data for the each condition, while the centres denote the mean values. For representational convenience, the relative changes in the distortion measure were scaled by the standard deviation across all conditions, for each measure.

Cross-correlation coefficients were computed between the μ measured for each utterance and the corresponding phoneme error, across 11 different noise conditions, including clean condition, for a total of 14,784 utterances (11*1344). These cross-correlation values are 0.77 for μ_{plp} and 0.84 for μ_{post}. For a comparison, results using the earlier proposed measure [4], denoted here as μ_{SD} are also shown.

As seen, while all measures increase with increasing noise levels, μ_{post} is the least sensitive to the type of the noise, and yields highest correlation with observed phoneme errors.

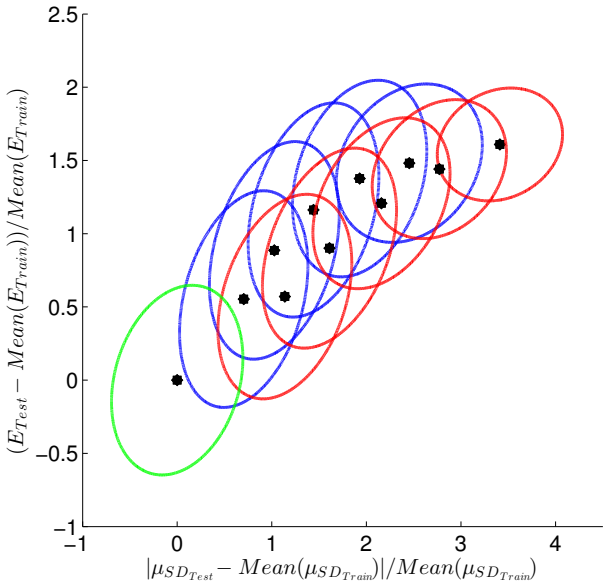


Fig. 3: Relative changes in the single-frame correlation based error prediction measure [4] *vs* sentence-level phoneme error for babble and buccaneer noises (NOISEX database) at SNRs 0dB, 5dB, 10 dB, 15 dB, 20dB

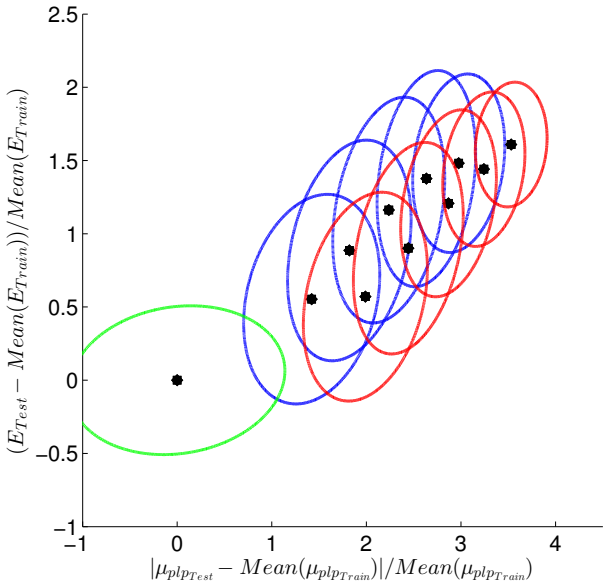


Fig. 4: Relative changes in the PLP-based error prediction measure S *vs* in sentence-level phoneme error for babble and buccaneer noises (NOISEX database) at SNRs 0dB, 5dB, 10 dB, 15 dB, 20dB

4.2. Effect of the time interval for the estimation

In order to use $M(\Delta t)$ in speech processing applications, it is desirable to derive a good estimate using a small portion of the speech segment. In order to understand the effect of the estimation interval on the estimate, TIMIT sentences were divided into groups of less than 2 s long (1452 sentences), 2-3 s long (6468 sentences), 3-4 s long (4609 sentences) and greater than 4 s (2255 sentences) and the

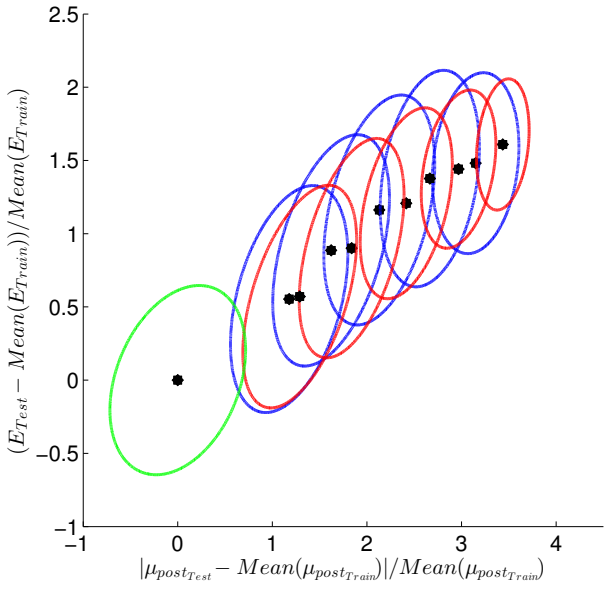


Fig. 5: Relative changes in the posterior-based error prediction measure S *vs* relative increase in sentence-level phoneme error for babble and buccaneer noises (NOISEX database) at SNRs 0dB, 5dB, 10 dB, 15 dB, 20dB

correlation coefficient was computed for each group. These correlations are shown in the bar plot in Fig. 6 As expected, the correlations increase with the length of the sentence.

The $M_{plp}(\Delta t)$ measure has the weakest correlation among the three measures compared here. However, it does not require estimation posterior probabilities of phonemes, is much easier to compute, and we assume it would be less sensitive to factors such as language, which influence the posterior estimation process. Further it can be readily used in applications where posteriors are not available.

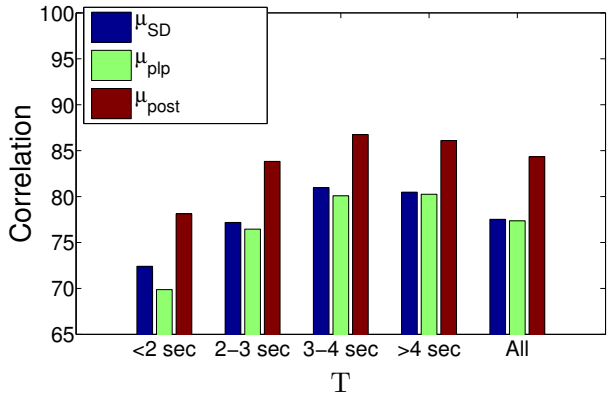


Fig. 6: Correlation between several performance prediction methods (Statistical Divergence [4], $M_{plp}(\Delta t)$ and $M_{post}(\Delta t)$ and recognition error for different of utterance lengths.

The posterior based prediction was applied for stream selection in a multi stream recognition system described elsewhere [9], where it has shown substantial improvement in recognition accuracy in phoneme recognition on noisy data. These automatically obtained results compared well with the best possible "cheating" results, where the most reliable streams were selected by a "human-in-the-loop", who knew the correct answers in advance.

5. DISCUSSION AND CONCLUSIONS

$M_{plp}(\Delta t)$ derived from PLP cepstra, peaks at about average phoneme length (around 70 ms between centers of phonemes) and stays approximately constant after about 200 ms. This observation supports the notion speech is composed of sequences of phonemes spaced roughly 70 ms apart, overlapping with their immediate neighbors. The 200 ms duration signifies extent after which the longest phoneme ceases to influence its neighboring phonemes. $M_{post}(\Delta t)$ computed from phoneme posterior features does not exhibit a peak at phoneme center time span because posterior features are trained to be constant over the phoneme lengths. Value of $M(\Delta t)$ for longer time-spans ($\Delta t > 200ms$) decreases with decreasing SNR as stationary distortions dominate the signal and feature vector values, making these vectors more similar. As a result, the divergence values among these feature vectors decrease. Thus $M(\Delta t)$ appears to be a good predictor of error rate of phoneme classifier that was trained on clean data.

6. REFERENCES

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

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

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

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

[5] H. Hermansky and J. Cohen, "Report from ws 1996," Tech. Rep., Center for Language and Speech Processing, the Johns Hopkins University, 1996.

[6] S. Badiezadegan and R. Rose, "A performance monitoring approach to fusing enhanced spectrogram channels in robust speech recognition," in *Proceedings Interspeech*, 2011, pp. 4780–4703.

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

[8] A. Varga, H.J.M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," Tech. Rep., DRA Speech Research Unit, Malvern, 1992.

[9] E. Variani, F. Li, and H. Hermansky, "Multistream recognition of noisy speech with performance monitoring," Tech. Rep., Johns Hopkins University, 2013.