



Intra-Frame Variability as a Predictor of Frame Classifiability

Trond Skogstad, Torbjørn Svendsen

Department of Electronics and Telecommunications
 Norwegian University of Science and Technology
 Trondheim, Norway

tronds@iet.ntnu.no, torbjorn@iet.ntnu.no

Abstract

This paper examines the association between the variability of the speech signal inside an analysis frame and the relative difficulty of classifying that frame. We introduce a novel measure of speech frame variability and show through classification experiments that this measure is a strong predictor of classifiability, even when conditioning on the distance to segment boundaries. Finally, we show how to incorporate the measure as weights in the discriminant function of a GMM-HMM recognizer, thereby increasing the relative importance of low variability frames in both decoding and training. This is shown to give a reduction in error rates.

Index Terms: stationarity, variability, estimation uncertainty

1. Introduction

The speech signal consists of phones of highly variable duration. Sometimes the signal is rapidly changing, never acoustically stable, at other times the signal is more or less stationary for hundred of milliseconds. Most current ASR systems attempt to describe this signal by a homogenous series of feature vectors based on short-time spectral analysis with fixed-length, equidistant measurements, modeled by continuous density hidden Markov models (CD-HMMs). The choice of frame length (usually 20-30 ms) is a compromise between the need for stationarity to ensure the validity of Fourier analysis and a desire for sufficient data samples to obtain reliable feature estimates. The time shift between frames (usually 10 ms) regulates the temporal resolution and should be chosen to capture the acoustic variability and, at least to some extent, to comply with the HMM assumption of state conditional independence. The feature extraction is thus designed to work well on average, occasionally violating the stationarity assumption, at other times using less samples for estimation than what is possible.

Recognizing the disparity between the rigid structure of traditional feature extraction and the fluidity of the speech signal, several attempts to modify the feature extraction process have been made. In [2] 13 different feature sets are extracted, using different combinations of time-frequency resolutions, spectral representations and temporal basis function to provide a diverse view of phonemic segments and segment boundaries. Each feature set serves as the input to a separate classifier. When these classifiers are combined through a committee-based approach, the classifying performance is among the best published for the TIMIT database. In [3] a new set of time-varying cepstral coefficients is introduced, where the stationarity assumption is replaced with the assumption that the coefficients are expressible in a known set of basis functions. Other authors suggest a two step approach, where the characteristics of the signal at

different times are examined in a pre-analysis step which then serves as a map for the final feature extraction. In [4] the focus is on detection of landmarks that signal significant articulatory changes. The related approach of detecting stationary segments is pursued in [5]. Here, hypothesis testing is done to find quasi-stationary segments that are as long as possible. This is shown to give a small, but significant, increase in recognition performance when compared to fixed scale analysis. This approach utilizes knowledge of signal variation to increase feature estimation reliability and thereby discrimination between classes. In a similar line of thought, this knowledge can be integrated to the back end recognizer rather than the front end feature extraction, that is, to keep the scale of analysis fixed and vary the relative importance of frames rather than keeping the importance fixed and vary the scale. This is the goal of the present work. To this end a measure of the variability inside an analysis frame (intra-frame variability) is introduced. We will seek to use this measure to quantify the uncertainty in the estimation of cepstral coefficients and relate this to error rates in frame classification. After establishing the relationship between frame variability and frame classifiability, we will show how this information can be utilized in a GMM-HMM framework, by integrating frame variability as weights in the discriminant function.

2. Estimator of intra-frame variability

Stationarity is an abstract concept, stringently defined, convenient to assume, but impossible to detect with certainty from a signal. A more tangible approach is to measure the variability of the features actually used in recognition, e.g., Mel-Frequency Cepstrum Coefficients (MFCCs).

The MFCCs $c[n]$ of a frame are calculated as the DCT of

$$S[m] = \ln\left(\sum_{k=0}^{N-1} |X_a[k]|^2 H_m[k]\right), \quad 0 < m \leq M \quad (1)$$

where X is the N -point DFT of the input signal and H_m is the m -th filter in a bank of filters uniformly spaced in the mel-scale [1]. As a first step in estimating the variability of the MFCCs, $c[n]$, each frame is divided into $J = \frac{N}{N_s}$ nonoverlapping subframes. For each subframe, j , a spectral estimate X_j is calculated by zero padding the waveform to N points and multiplying with \sqrt{J} before performing the DFT. The MFCCs $c_j[n]$ of the subframes are then obtained in normal fashion, using (1). A measure of the intra-frame variability can then be defined as

$$v = \frac{1}{J} \sum_{n=1}^M \sum_{j=1}^J (c_j[n] - c[n])^2. \quad (2)$$

This measure is in the form of a sum of sample variances, although $c[n]$ is not a true mean value of the $c_j[n]$ s. The zeroth MFCC, representing frame energy, is omitted in (2), as the term is substantially larger than the others and would dominate the sum. The MFCCs could also be further filtered to compensate for the difference in variance of the cepstral coefficients, although caution should be exercised when placing increased importance on the high order coefficients as these can be spurious when estimated over few samples.

As the periodogram spectral estimator is the convolution of the true power spectral density with the Fourier transform of a Bartlett window of length $2N$, the periodogram is a biased estimator of the true PSD for finite N [6], with increasing bias for decreasing N . A result of this is that the $c_j[n]$ s are biased estimators of $c[n]$. If J is kept reasonably small this problem is of no significant practical consequence.

3. Large margin models

The large margin GMM-HMM framework [8] was used for all classification and recognition experiments in this work. This framework combines the discriminative training criterion used in support vector machines with the usual GMM-HMM models suited to multiway classification of sequential data. The software used was obtained from the author of [8], and this short review of the framework will follow roughly his treatise. Note that all constraints are given without the slack variables representing margin violations. In optimization the sum of margin violations over the training set is balanced with a regularization term and minimized.

In multiway classification using GMMs, the maximum a posteriori decision rule can be expressed as

$$y = \arg \min_c \{(\mathbf{x} - \mu_c)^T \Sigma_c^{-1} (\mathbf{x} - \mu_c) + \theta_c\}, \quad (3)$$

where \mathbf{x} is the observation vector, μ_c is the class mean, Σ_c^{-1} is the inverse class covariance matrix and θ_c is a nonnegative scalar offset related to the class prior probability. By collecting the parameters $\{\mu_c, \Sigma_c^{-1}, \theta_c\}$ as

$$\Phi_c = \begin{bmatrix} \Sigma_c^{-1} & -\Sigma_c^{-1} \mu_c \\ -\mu_c^T \Sigma_c^{-1} & \mu_c^T \Sigma_c^{-1} \mu_c + \theta_c \end{bmatrix} \quad (4)$$

the decision rule simplifies to

$$y = \arg \min_c \{\mathbf{z}^T \Phi_c \mathbf{z}\}, \text{ where } \mathbf{z} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}. \quad (5)$$

Large margin training of GMM classifiers seeks to separate each labeled example with at least one unit distance from the decision boundary of all competing classes. That is, for a single frame classifier,

$$\mathbf{z}_t^T (\Phi_c - \Phi_{y_t}) \mathbf{z}_t \geq 1, \quad \forall c \neq y_t. \quad (6)$$

To represent each class by multiple mixture components each example \mathbf{z}_t can be given a proxy mixture label m_t as well as a class label y_t . This can be done by fitting an M component GMM for each class using maximum likelihood estimation and then labeling each example with the mixture component giving the highest posterior probability. By using the softmax inequality $-\log_m e^{-a_m} \leq \min_m a_m$ eq. (6) can be extended to separate each example from all components of competing classes without increasing the number of constraints:

$$-\log \sum_m e^{-\mathbf{z}_t^T \Phi_{c m} \mathbf{z}_t} - \mathbf{z}_t^T \Phi_{y_t m_t} \mathbf{z}_t \geq 1, \quad \forall c \neq y_t. \quad (7)$$

The large margin framework can also be extended to CD-HMM models for sequential classification in ASR. In this case, the discriminant function is the usual

$$\mathcal{D}(\mathbf{X}, \mathbf{y}) = \sum_t [\lambda(y_{t-1}, y_t) + \rho(\mathbf{x}_t, y_t)] \quad (8)$$

where \mathbf{X} is the observation sequence, \mathbf{y} the state sequence, $\lambda(y_{t-1}, y_t)$ the transition scores and $\rho(\mathbf{x}_t, y_t)$ the emission scores, parameterized by large margin GMMs:

$$\rho(\mathbf{x}_t, y_t) = \log \sum_m e^{-\mathbf{z}_t^T \Phi_{y_t m} \mathbf{z}_t} \quad (9)$$

To handle sequences of different lengths and disparate divergence from the target sequence, a natural extension of the large margin principle is to try to separate all competing sequences with at least the Hamming distance (number of mismatched labels) from the target sequence:

$$\mathcal{D}(\mathbf{X}, \mathbf{y}^*) - \mathcal{D}(\mathbf{X}, \mathbf{y}) \geq h(\mathbf{y}, \mathbf{y}^*), \quad \forall \mathbf{y} \neq \mathbf{y}^*. \quad (10)$$

The number of constraints is exponential in the sequence length T , but can be folded to a single constraint using the softmax inequality once more:

$$-\mathcal{D}(\mathbf{X}, \mathbf{y}^*) + \log \sum_{\mathbf{y} \neq \mathbf{y}^*} e^{h(\mathbf{y}, \mathbf{y}^*) + \mathcal{D}(\mathbf{X}, \mathbf{y})} \leq 0. \quad (11)$$

To keep the optimization convex, the discriminant function for the target sequence is often restricted to a single mixture component m_t^* per state y_t^* . This is done using proxy labels as in eq. (7).

3.1. Variability weighting of emission scores

Supposing a connection between frame variability and frame classifiability can be established, how can this knowledge source be integrated to the GMM-HMM recognizer framework? Consider eq. (3). The (squared) Mahalanobis distance between the feature vector \mathbf{x}_t and class centroid \mathbf{u}_c implemented in the decision rule generalizes the (squared) Euclidian distance to account for the variance in the direction of the difference vector $\mathbf{x}_t - \mathbf{u}_c$. In the context of sequential classification the accumulated emission score is the sum of the (squared) Mahalanobis distances between the feature vector sequence \mathbf{X} and the hypothesized sequence of class centroids. If the feature estimation uncertainty of frame x_t is assumed to be centered on the feature vector x_t and spherical in the feature vector components with a variance of σ_{et}^2 , an additional, Mahalanobis-like, weighting of the emission scores according to their feature estimation uncertainty can be implemented by changing the discriminant function in eq. (8) to

$$\mathcal{D}(\mathbf{X}, \mathbf{y}) = \sum_t \left[\lambda(y_{t-1}, y_t) + \frac{1}{\sigma_{et}^2} \rho(\mathbf{x}_t, y_t) \right]. \quad (12)$$

In this work the variance σ_{et}^2 is assumed to be measurable from the waveform by the methods in sec. 2 and related to the frame variability v_t as $\sigma_{et}^2 = v_t + v_{floor}$, where v_{floor} is a constant acting as a lower bound on variability to avoid excessively large weights. When comparing competing sequence hypotheses using the discriminant function of eq. (12), an increased importance is placed on the frames in the sequence of comparatively low variability. Further, during training, stricter margin

constraints are enforced on segments of comparatively low variability, or equivalently: margin violations are punished more severely when occurring in segments of comparatively low variability.

4. Experiments

The experimental section is divided in three subsections. The first part serves to establish the relationship between frame variability and 1) phonetic category and 2) distance to the boundaries of the manual phonetic labeling. The following subsection covers the frame level classification experiments, which will demonstrate the significance of intra-frame variability for classification performance. Here, the 13 first MFCCs (including the zeroth coefficient) were used as features. Δ - and $\Delta\Delta$ -coefficients were not included in the feature vector for this experiment as they contain no information on the current frame. The final experiments are on the tasks of continuous phonetic recognition. Here the goal was to demonstrate that variability weighting of emission scores can be beneficially included in a realistic, high performance recognition system. Hence, the standard full 39 component MFCC feature vector was used. In all experiments the frame length used in estimating $c[n]$ was kept at 25 ms and the frame shift was 10 ms. Each frame was divided into $J = 5$ subframes. Rectangular windows were used for both frames and subframes in the variability calculations, while a Hamming window was used in calculations of the MFCCs used as features in the recognition experiments.

All experiments were conducted on the TIMIT acoustic-phonetic corpus. The utterances in the TIMIT database are phonetically labeled and segmented into nonoverlapping regions. We used the standard NIST 462 speaker training set and the 24 speaker core test set. A 50 speaker development set was also used, overlapping neither with the training set nor the test set. In the classification and recognition experiments, following common practice [7], the 61 phonetic labels were folded into 48 phones before training the classifiers. Before evaluation, these 48 phones were further folded into 39 categories.

4.1. Characteristics of high variability frames

To investigate the relationship between a frame’s variability and its phonetic label, we divided the 61 labels into the seven phonetic categories vowels, plosives, affricates, fricatives, nasals, semivowels and silence in accordance with the TIMIT documentation [9]. The intra-frame variability v of each frame in the training set was calculated by (2). After sorting the frames in ascending order of v , five subsets ($S_{10\%}, S_{20\%}, \dots, S_{50\%}$) were created, containing respectively the 10%, 20%, \dots , 50% frames of highest variability. For each set, the relative representation of each phonetic category was calculated as (number of frames in phonetic category in set)/(number of frames in phonetic category in training set). In Fig. 1 plosives, semivowels and vowels can be seen to be overrepresented in the high variability sets, the nasals and affricates seem to be proportionately represented, while fricative and silence frames are less frequent in the high variability sets than in the whole training set. In the classification experiments of section 4.2 comparing sets of high variability frames with the set of all frames, the difference in the distribution of phonetic categories could present a bias. In particular, the relative lack of silence frames could inflate the error rates for the sets of high variability somewhat.

Intuitively there should be a strong connection between the variability of a frame and the distance to a segment boundary,

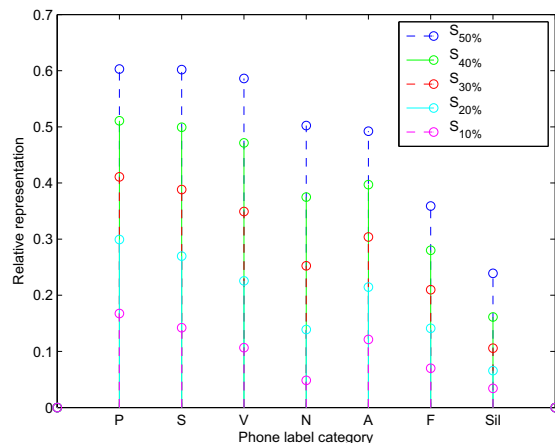


Figure 1: Representation of different phone label categories in the top deciles of intra-frame variability, relative to total number of frames in category. Phonetic categories: P: Plosives, S: Semivowels, V: Vowels, N: Nasals, A: Affricates, F: Fricatives and Sil: Silence.

typically placed at the peak point of transition between two phonetic segments. Fig. 2 demonstrates that this is indeed the case. While frames neighboring a segment boundary accounts for 25% of the training set, it accounts for 50% of the frames in the set $S_{10\%}$. For larger distances the frames are distributed in a similar fashion in all sets. While this is a satisfactory result, verifying a common sense assumption of the nature of variability, it also presents a problem: the comparison of classification experiments across different sets with different distributions is not straightforward. Phonetic labeling of transitional areas is not easy, and when an exact point of transition between two segments is impossible to find or even define, an arbitrary choice has to be made. Consequently, the labels of boundary frames contain an inherent uncertainty that could translate to increased error rates for sets where boundary frames are represented in proportions higher than average.

4.2. Frame classification experiments

To separate the effect of indistinct phonetic labeling from the influence of intra-frame variability on classifiability we perform frame classification experiments conditioned on a minimum distance from the boundaries d_{min} as well as frame variability v . This is depicted in Fig. 3. The test data frames’ set membership were determined from their estimated variability v_t and thresholds calculated on the training set. The bottom line is for the entire test set, the others for the sets of high variability frames. The horizontal axis shows d_{min} , where a value of for instance $d_{min} = 2$ means that all frames residing less than 2 frames from a boundary are excluded. The gap in error rates between the entire test set and the sets of high variability can be seen to be high, e.g., over 10 percent absolute for the set $S'_{30\%}$, and consistent across all d_{min} - actually slightly increasing for increasing d_{min} . The slight tilt in the uppermost curve is probably due to few samples in the test set with both high d_{min} and high v . The error rate can be seen to be strictly increasing for increasing v . Hence, it is fair to conclude that intra-frame variability as defined in (2) is a strong predictor of frame classifiability.

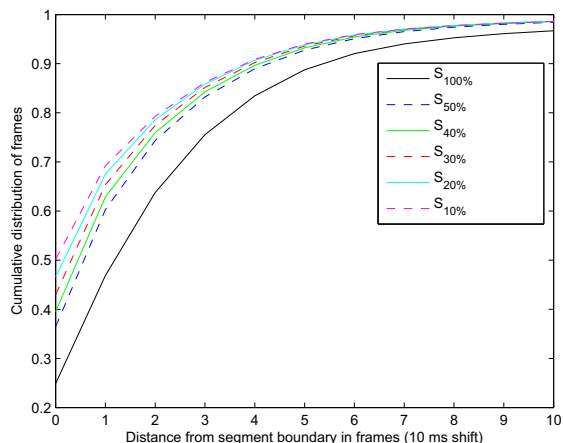


Figure 2: Cumulative distribution of distance from segment boundary, measured in frames, for the sets of high frame variability. The black line ($S_{100\%}$) is for the entire training set.

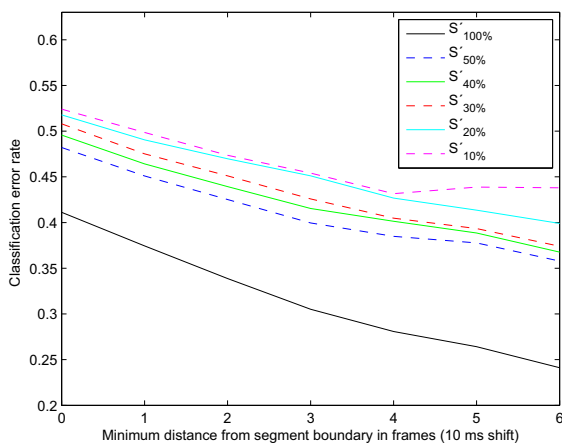


Figure 3: Classification error rates for frames conditioned on both variability and minimum distance to segment boundary, d_{min} , rendered on the horizontal axis.

4.3. Recognition experiment

To demonstrate the appropriateness of accounting for frame estimation uncertainty through a variability weighting scheme, we performed continuous recognition experiments comparing a baseline (BL) system of a large margin trained GMM-HMM as in [8] with a variability weighted (VW) system using the discriminant function eq. (12). The phone error rates in percents are given in table 1. For completeness, the results are reported on both the development set and the core test set. The development set is used to determine a language/acoustic score scale parameter and the regularization factor, but this is done on both the baseline and the proposed system, so the comparison is fair. For the systems using 1-4 mixture components, the results for both BL and VW are slightly better than the results reported in [8]. In the case of 8 mixture components per state both BL and VW have slightly higher error rates compared to [8], perhaps suggesting a need of further optimization of the regularization factors. The proposed system can be seen to give a performance

gain compared to the baseline system in all test cases but one (1 mixture component, tested on the development set), and although the differences between system performance are probably not statistically significant at conventional levels in any one test case, the experiments as a whole should indicate the usefulness of the proposed method.

mixture components	BL dev	VW dev	BL test	VW test
1	28.8	29.0	30.9	30.7
2	28.1	27.9	30.5	29.9
4	28.0	26.9	29.7	29.2
8	27.8	26.8	29.6	28.8

Table 1: Phone error rates in percent for continuous phone recognition experiment.

5. Conclusions

This paper examines the association between the variability of the speech signal inside an analysis frame and the relative difficulty of classifying that frame. To this end we introduced a novel measure of intra-frame variability and explored the relationship between variability and different phonetic categories as well as variability and distance to segment boundaries. In frame level classification experiments performed on the TIMIT corpus we demonstrated the strong connection between high variability and high classification error rates. After giving a short review of the large margin GMM-HMM we showed how intra-frame variability could be used as weights on the emission scores in sequential classification, increasing the importance of low variability frames in the decoding process and enforcing stricter margin constraints on segments of comparatively low variability. Finally, the proposed method is shown to lower error rates in a continuous phone recognition experiment.

6. References

- [1] Huang, X., Acero, A., Hon, H-W., “Spoken Language Processing”, Prentice Hall PTR, New Jersey, 2001.
- [2] Halberstadt, A.K., and Glass, J.R., “Heterogeneous measurements for phonetic classification”, Proc. of European Conf. Speech Communication and Technology, 1997.
- [3] Skogstad, T., and Svendsen, T., “Time-varying Cepstral Coefficients”, Proc. of ISCA ITRW on Speech Analysis and Processing for Knowledge Discovery, 2008.
- [4] Espy-Wilson, C.Y., Pruthi, T., Juneja, A., Deshmukh, O., “Landmark-based Approach to Speech Recognition: An alternative to HMMs”, Proc. of Interspeech 2007, Antwerpen, pp. 886-889
- [5] Tyagi, V., Boulard, H., Wellekens, C., “On variable-scale Piecewise Stationary Spectral Analysis of Speech Signals for ASR”, Speech communication, Vol. 48, Issue 9. September 2006, pp. 1182-1191
- [6] Kay, S.M. “Modern Spectral Estimation, Theory & Application”, Prentice Hall PTR, New Jersey, 1987.
- [7] Lee, K., Hon. H., “Speaker-independent phone recognition using hidden Markov models”, IEEE Trans. Acoust., Speech, and Signal Processing, Vol. 37, No. 11, 1989, pp. 1641-1648
- [8] Sha, F., “Large Margin Training of Acoustic Models for Speech Recognition”, Ph.D thesis University of Pennsylvania, 2007
- [9] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S. and Dahlgren, N. L., “DARTPA TIMIT Acoustic-Phonetic Continuous Speech Corpus”, U.S. Dept. of Commerce, NIST, Gaithersburg, MD, February 1993