INTERSPEECH 2009 BRIGHTON

Discriminant Spectrotemporal Features for Phoneme Recognition

Nima Mesgarani[1], *G.S.V.S. Sivaram*[1,2], *Sridhar Krishna Nemala*[1], *Mounya Elhilali*[1], *Hynek Hermansky*[1,2]

[1] The Center for Language and Speech Processing
[2] Human Language Technology, Center of Excellence
Johns Hopkins University
nmesgar1,sivaram,nemala,mounya,hynek@jhu.edu

10.21437/Interspeech.2009-755

Abstract

We propose discriminant methods for deriving two-dimensional spectrotemporal features for phoneme recognition that are estimated to maximize the separation between the representations of phoneme classes. The linearity of the filters results in their intuitive interpretation enabling us to investigate the working principles of the system and to improve its performance by locating the sources of error. Two methods for the estimation of filters are proposed: Regularized Least Square (RLS) and Modified Linear Discriminant Analysis (MLDA). Both methods reach a comparable improvement over the baseline condition demonstrating the advantage of the discriminant spectrotemporal filters.

Index Terms: phoneme recognition, spectrotemporal filters, data driven features

1. Introduction

Automatic Speech Recognition can benefit from using features that do not merely reflect the short-term spectral profile of speech but are derived from longer temporal windows [1]. These features can have a fixed temporal resolution [1], or perform a multiresolution temporal decomposition on each frequency band of the signal [2]. The long-term features have also been generalized to 2D spectrotemporal patterns [3,4]. One motivation that justifies this idea is the known characteristics of the neurons in higher order auditory areas that are responsive to a wide range of temporal and spectral modulations of the auditory stimulus [5]. Many of the so-called Spectro-Temporal Receptive Fields (STRF) found in the auditory cortex show complex spectrotemporal characteristics suggesting their ability to perform complex filtering on the input time-frequency representation. In addition, recent findings by neurophysiologists have shown that the cortical neurons are not just static filters with fixed transfer functions [6]. In fact, the receptive field of cortical neurons changes actively when the brain engages in audio classification task to assist the discrimination of the sound classes [7]. A recent study has proposed a discriminant model generating spectrotemporal filters that closely match the observed neural receptive field changes in a variety of audio classification tasks [8]. The model in [8] assumes that the changes in receptive fields result in increased separability between the representations of sound classes.

In this paper, we propose similar discriminant algorithms as in [8] that optimize 2D spectrotemporal filters for discrimination of each phoneme from the rest. The outputs of such filters are used as inputs to a neural network trained to generate the phoneme posterior probabilities. The proposed 2D filters operate on the time-frequency representation of sound in a similar way as in [3,4]. In contrast, our spectrotemporal filters are found by maximizing phoneme separability in their projected space, not by *selecting* from a fixed set of parametric filters similar to [4].

We utilized two methods for finding the discriminant filters: (1) Regularized Least Square technique (RLS) [9] and Modified Linear Discriminant Analysis (MLDA) [10]. Since the spectrotemporal filters that are optimized to discriminate one phoneme from the rest are linear, their interpretation becomes very straightforward. This enables us to investigate the working principles of the system and to gain insights that can be used to improve the performance and to determine the sources of error. Next, we explain how the filters are derived and describe their intuitive operation.

2. Spectrotemporal filter estimation

We used two methods to estimate the discriminant spectrotemporal filters that were applied to the log critical band energies: RLS and MLDA. The optimization criteria for the two techniques is different, however they result in comparable performance.

2.1. Regularized Least Square algorithm

Regularized least square algorithm solves the following optimization problem [9]:

$$\frac{1}{2}\| Y - Xw \|^2 + \frac{\lambda}{2}\| w \|^2 \qquad (1)$$

where X is the data set, Y is the desired labels and w is the weight vector that minimizes equation 1. λ is a regularization constant controlling the tradeoff between fitting the training set accurately and finding a function with small norm (smoothness of the weighting function w). Since this is a differentiable convex optimization problem, the solution can be found to be

$$w = (X^T X + \lambda I)^{-1} X^T Y \qquad (2)$$

The parameter λ in the solution can be estimated by minimizing the leave-one-out (LOO) error. Due to the properties of linear regularized least square method, the computational cost of searching for the regularization parameter is very low, which means that searching for a good λ has about the same computational cost as solving the problem for a single λ [11]. This method results in one 2D spectrotemporal filter for each phoneme, which is the hyperplane that separates the instances of that phoneme from the competitors. Since the optimized filters are linear, we can plot them in exactly the same way as we plot the input time-frequency representation to investigate the spectrotemporal

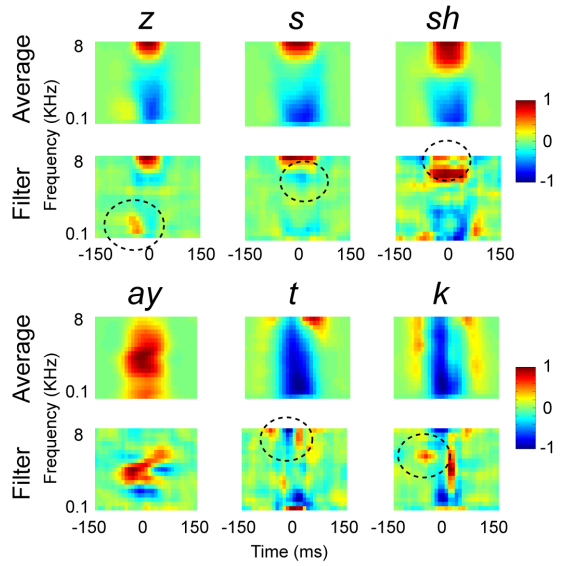

Figure 1. Average phoneme spectrograms (top rows) and optimized spectrotemporal filters using RLS algorithm for 6 phonemes. The filters are tuned to the discriminant features of the phonemes, for example the presence of mid-frequency energy for /sh/ that results in positive weight for /sh/ but negative for /s/ (highlighted by circles). The mask for /ay/ shows sensitivity to the upward and downward formant transitions of this vowel.

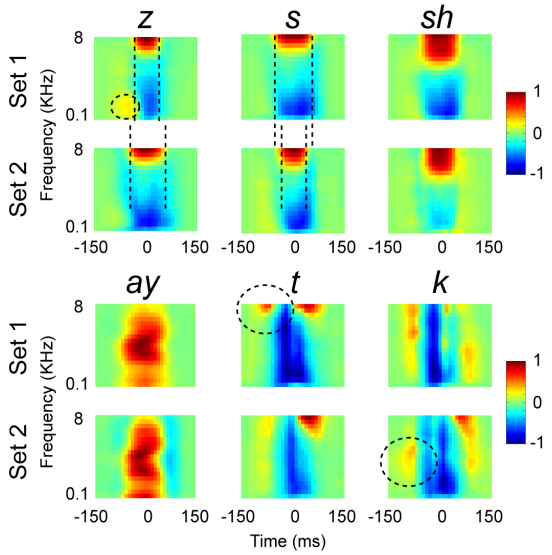

Figure 2. Average phoneme spectrograms of training samples that were correctly classified by the linear spectrotemporal filters in the first set (top rows) and second set (bottom rows). The average spectrograms display a systematic variability in the production of each phoneme. For example, exemplars of /s/ in the first set are longer than the ones in the second class (dashed lines).

features of phonemes that are used by the filters to discriminate them properly. For example, figure 1 displays the average time-frequency representation and the optimized 2D filter using RLS algorithm for three fricatives /z/, /s/ and /sh/, vowel /ay/ (as in *bite*) and two plosives /t/ and /k/. The average spectrograms of different fricatives are known to be closely related to their place of articulation [12]. For instance, the difference between the more forward places of articulation for /s/ compared to /sh/ is mirrored by the downward shift of the highpass spectral edge (Fig. 1). However, the optimal 2D spectrotemporal filters for the discrimination of these two phonemes reflect the differences between them. Since both /s/ and /sh/ produce high frequency energy but only /sh/ produces mid frequencies, the filters show negative mid-frequency weight for /s/, but positive weight for /sh/ as circled in Figure 1. The 2D filter for fricative /z/ shows a strong low frequency positive weight, signifying the presence of low-frequency energy during the production of /z/ which is due to its manner of articulation (voicing) [12]. For plosive /t/, the 2D filter is selective to a sudden change of energy in high frequency channels, compare to the /k/ discriminator which detects a change of energy in mid frequencies (as circled in figure 1). Finally, the 2D filter optimized for /ay/ is tuned to a spectrotemporal sweep, mostly to upward moving energy reflecting the formant transitions in the production of /ay/. This final example emphasizes the importance of spectrotemporal features for detecting complex patterns that can be difficult to detect by spectral or temporal only features.

The one versus rest spectrotemporal filters described above provide 39 spectrotemporal filters each optimized for discrimination of one phoneme in standard TIMIT phoneme set. To expand this idea, we repeated this optimization several times, each time using only the portion of the training set that was not correctly labeled by the 2D linear filter from the previous step. This idea is similar to boosting. However, instead of weighting the training samples, we used different subsets of the training set. In practice, this was done by excluding the training exemplars that were correctly labeled by the *linear* masks first, and then optimizing a new set of 2D filters for the remaining samples. This in effect captures the systematic variability that exists in the production of a particular phoneme as shown in Figure 2. This figure shows the average time-frequency representation of the same 6 phonemes as in figure 1, broken into two subsets that were correctly classified by the first set of discriminant filters (top rows) and the second set (bottom rows). Figure 2 shows substantial differences between the average phoneme spectrograms of the two groups. For example, the average duration of the high-frequency energy in the production of /s/ phonemes in the first group (54 ms) is longer than in the second one (36 ms), an effect that can only be captured by masks that span long temporal windows. For /z/, the first set shows a shorter high frequency burst in addition to stronger low frequency energy. Plosives /k/ and /t/ in two groups also display a different average spectral shape which may reflect the contextual effects. For example, the /t/ phonemes in the first group have more high frequency energy before the stop. The percentage of the training set that can be classified correctly using the 2D linear masks is shown in figure 3 as a function of filter numbers.

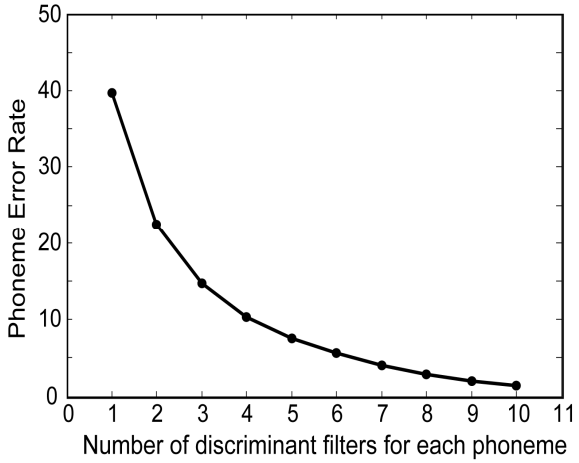

Figure 3. Phoneme error rate on the training examples vs. the number of spectrotemporal filter sets used. Each set is optimized for the classification of the samples that were not correctly recognized by the previous sets.

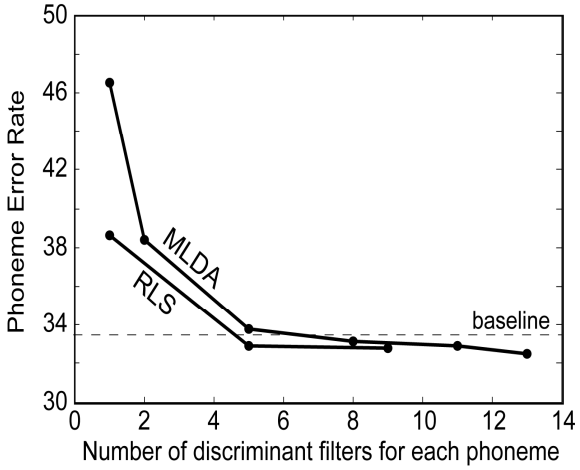

Figure 4. Phoneme error rate vs. number of feature sets obtained from MLDA and RLS algorithms. The system from RLS reaches a better performance for the same number of features; they both converge to about the same performance.

2.2. Modified linear Discriminant Analysis

Fisher Linear Discriminant Analysis (FLDA) is a standard technique for discriminant analysis and dimensionality reduction in pattern classification [13]. FLDA can be used to project the original high-dimensional data onto a low-dimensional space where the classes are well separated by maximizing the ratio of between-class (S_b) to within-class scatter matrix (S_w):

$$J(\phi) = \frac{trace(\phi^T S_b \phi)}{trace(\phi^T S_w \phi)} \quad (3)$$

Due to the rank limitation of S_b, for a two-class problem, we can only get one optimal discriminating vector. This shortcoming, when used for dimensionality reduction, limits us to a low-dimensional space of one and thus may hinder classification performance. Modified Linear Discriminant Analysis (MLDA) [10] is a generalization of FLDA that overcomes this limitation. MLDA uses the same optimization criteria as FLDA, however the definition of the between-class scatter matrix (S_b) is modified as follows:

$$S_b = \sum_{p=1}^{c} \frac{N_p}{N} \sum_{j=1}^{N_p} \sum_{k=1, k \notin I_p}^{N} (x_{pj} - x_k)(x_{pj} - x_k)^T$$

where c is the number of classes, N is the total number of training samples, N_p is the number of samples in the p_{th} class C_p , x_{pj} is the j_{th} training sample in the p_{th} class, x_k is the k_{th} sample in the training set and I_p is the index set of class C_p [10]. We can vary the number of discriminant projections (*d*) by choosing the first *d* Eigen vectors of the following equation corresponding to the *d* largest Eigen values:

$$S_w^{-1} S_b \phi_d = \phi_d \Lambda_d \quad (5)$$

where Λ_d contains the Eigen values. Using this technique, we found multiple projections for the discrimination of each phoneme from the rest.

3. Experiments

We conducted speaker independent phoneme recognition experiments on TIMIT to test the effectiveness of the proposed features. 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 data set consists of 1344 utterances from 168 speakers. The power spectrum of the speech signal was estimated 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 are then estimated from the power spectrum using bark frequency weights. The features are then obtained by convolving the 2D log-critical band representation with the proposed discriminant filters. These projected log critical band energies onto the spectrotemporal filters were used to train a Multi-layer perceptron (MLP) with a single hidden layer to convert input features into posterior probabilities of phonemes (standard set of 39). These probabilities are converted to scaled likelihoods, which are then decoded by applying Viterbi algorithm with a minimum duration of three states per phoneme.

In all of our experiments, the number of hidden nodes of MLP is chosen such that the total number of free parameters remains constant to minimize the effect of classifier complexity. The performance of a feature set is evaluated using Phoneme Error Rate (PER). We used Asymmetric MRASTA (AMRASTA) [14] as our baseline. These recently developed features are an alternative to MRASTA [2] features that are obtained from a temporal trajectory of the log-critical band energies by filtering it using a bank of multiresolution filters. The difference between AMRASTA and MRASTA is the asymmetric shape of the impulse response of AMRASTA which results in their improved performance [14]. The performance of the baseline and two proposed spectrotemporal

features are shown in Table I. Also, Table I includes the performance of the system trained using 9 frames of PLP which include delta and double delta features (dimensionality = 351). The discriminant features from RLS (10 set of 39 spectrotemporal filters, dimension = 390) and MLDA (13 set of 39 spectrotemporal filters, dimension = 507) perform well compared to the AMRASTA (dimension = 504) features. Both these algorithms show an absolute improvement of about 1.0% and 1.3% respectively, over the baseline.

To study how the number of spectrotemporal discriminant filters affects the classification accuracy, we conducted phoneme recognition tests using variable number of filter sets (each set contains 39 spectrotemporal filters optimized for the 39 phonemes). For RLS, the number of filters is increased by eliminating the correctly classified training samples and estimating a new set of filters on the remaining training samples (section 2). For the MLDA filters, we changed the number of projections for each phoneme as described in section 2.

Figure 4 summarizes the effect of number of filters on the recognition accuracy of the test set. The performance of both systems improves as the number of discriminant filters increases. Figure 4 also demonstrate that for the same number of features, the discriminant features from RLS algorithm outperform the ones from MLDA algorithm, suggesting a better separation between the projections of phoneme classes for RLS filters.

AMRASTA	RLS filters	MLDA filters	9 frame PLP
33.8	32.8	32.5	33.1

Table I. PER for the baseline (AMRASTA), spectrotemporal features (RLS and MLDA) and 9 frame PLP.

4. Discussion

We proposed a spectrotemporal feature extraction method based on optimized discriminability of phoneme representations. The features increased the performance of system compared to the baseline condition. Each filter is estimated to discriminate one designated phoneme from the rest and by repeating this procedure over the portions of the misclassified training data we derive multiple filters for each phoneme. The various filters presumably capture the variability in the production of each phoneme due to context and speaker effects.

There are several directions that in the future can be explored. First, motivated by results of psychoacoustic experiments [15], it may be beneficial if each filter is constrained to a limited frequency region of the input spectrogram, which may particularly be important for noise robustness. To achieve this goal, several filters can be estimated for each phoneme by limiting the optimization to certain frequency bands, or temporal windows. The performance of the spectrotemporal features in noise remains an issue of study. In the current work, we estimated the 2D filters considering only the phoneme templates that are centered at the middle of the phoneme. To achieve translational invariance, it may be useful if adjacent time-frequency patterns are also included in the estimation of spectrotemporal filters. In addition, we would like to explore the possibility of adding spectrotemporal filters that target certain phoneme confusions as a way to improve the accuracy of the system.

5. Acknowledgement

Partial funding for this project was obtained from the Air Force Office of Scientific Research, and the National Institutes of Health (NIH) Grants R01DC005779.

6. References

[1] Hermansky, H., Morgan, N., “RASTA processing of speech”, IEEE Trans. on Speech and Audio Processing, vol. 2, num. 4, pp. 578-589, Oct, 1984

[2] Hermansky, H., Fousek, P., “Multi-resolution RASTA filtering for TANDEM-based ASR”, Interspeech, pp. 361-364, Sep. 2005

[3] Kleinschmidt, M., Gelbart, D., “Improving Word Accuracy with Gabor Feature Extraction,”*Proc. Of ICSLP*, Colorado, USA, 2002

[4] Meyer, B., Kollmeier, B., “Optimization and evaluation of Gabor feature sets for ASR”, Proc. Interspeech 2008

[5] Klein D. J, Depireux D. A., Simon J. Z., and Shamma S. A.,”Robust spectrotemporal reverse correlation for the auditory system: optimizing stimulus design”, *J Comput Neurosci* 9: 85-111, 2000

[6] Fritz J. B., Shamma S. A., Elhilali M., and Klein D., “Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex”, *Nat Neurosci* 6: 1216-1223, 2003

[7] Fritz J. B., Elhilali M., and Shamma S. A., ”Adaptive changes in cortical receptive fields induced by attention to complex sounds. *J Neurophysiol* 98: 2337-2346, 2007

[8] Mesgarani, N., Fritz, J. B., Shamma, S. A., "A computational model of rapid task-related plasticity of auditory cortical receptive fields", to appear in Journal of Computational Neuroscience, 2009

[9] Rifkin, R., Mesgarani, N., "Discriminating speech and non-speech with regularized least squares", Interspeech, paper 1779, 2006

[10] Chen, S. C. and Li, D.H., ”Modified Linear Discriminant Analysis”, Pattern Recognition, vol. 38, pp. 441-443, 2005

[11] Rifkin, R., “Everything Old Is New Again: A Fresh Look at Historical Approaches to Machine Learning”, Ph.D. thesis, Massachusetts, Institute of Technology, 2002

[12] Ladefoged P. “Course in Phonetics”, Heinle, 2005

[13] Mclachlan, G. J., “Discriminant Analysis and Statistical Pattern Recognition”, Wiley, New York, 1992

[14] Sivaram, G.S.V.S., Hermansky, H., ``Introducing temporal asymmetries in feature extraction for automatic speech recognition", Interspeech, Brisbane, 2008

[15] Lippmann, R., “Accurate consonant perception without mid-frequency speech energy”, IEEE Trans. Speech and Audio Proc., Vol. 4, No. 1, 1996