INTERSPEECH 2009 BRIGHTON

Discriminative Feature Transformation using Output Coding for Speech Recognition

Omid Dehzangi[1], *Bin Ma*[2], *Eng Siong Chng*[1], *Haizhou Li*[1,2]

[1] School of Computer Engineering, Nanyang Technological University, Singapore
[2] Institute for Infocomm Research, Singapore

{dehzangi, aseschng}@pmail.ntu.edu.sg, {mabin, hli}@i2r.a-star.edu.sg

10.21437/Interspeech.2009-754

Abstract

In this paper, we present a new mechanism to extract discriminative acoustic features for speech recognition using continuous output coding (COC) based feature transformation. Our proposed method first expands the short-time spectral features into a higher dimensional feature space to improve its discriminative capability. The expansion is performed by employing the polynomial expansion. The high dimension features are then projected into lower dimension space using continuous output coding technique implemented by a set of linear SVMs. The resulting feature vectors are designed to encode the difference between phones. The generated features are shown to be more discriminative than MFCCs and experimental results on both TIMIT and NTIMIT corpus showed better phone recognition accuracy with the proposed features.

Index Terms: speech recognition, discriminative features, polynomial expansion, output coding, SVM

1. Introduction

The conventional approach for automatic speech recognition (ASR) is to use the short-time spectral features, such as mel-frequency cepstrum coefficients (MFCCs), as the acoustic feature vectors, and a set of Gaussian mixture based hidden Markov models (HMMs) as the acoustic models. Model parameters in such acoustic models are normally estimated using the maximum likelihood (ML) criterion. In the past decade, discriminative training techniques, as opposed to ML modeling, were extensively studied for improved ASR performance [1,2,3,4].

As the input to ASR systems, conventional acoustic feature vectors, which carry spectral information of the speech signal, are not designed by optimizing a discriminative measure. In recent years, there has been much research interest to improve the discriminative capability of the features by applying linear/nonlinear discriminative transformation on the original features [5,6,7]. E.g, the Linear Discriminant Analysis (LDA) has been shown to improve discrimination in the speech feature space and led to improve recognition performance [5]. An example of nonlinear transformation is the hybrid connectionist-HMM systems [6] approach which uses discriminatively-trained neural networks to estimate the probability distribution among sub-word units given the acoustic observations. In another effort, TANDEM connectionist feature extraction [7] combines neural-net discriminative feature processing with Gaussian-mixture distribution modeling. Inspired by discriminative training of acoustic models, recent research like in feature-space MPE [8] and MMI-splice [9] showed that using discriminative criteria in optimizing feature projection function is effective to improve speech recognition accuracy.

In this paper, we propose a new mechanism to extract features for speech recognition using discriminative feature transformation. A high dimensional feature space for better discriminative capability is constructed by polynomial expansion of the original acoustic features. The expanded features are then passed to a set of linear discriminants that projects the sequences of expanded vectors into a lower dimensional space with the objective to discriminate among the phone classes using output coding technique [10]. Differently from LDA, the extracted features benefits from a distribution free method which is suitable in the high dimensional space. As opposed to TANDEM which is a nonlinear transformation using multilayer perceptron (MLP) neural networks, we adopt linear SVMs to transform the high dimensional feature vectors for computational efficiency.

The rest of this paper is organized as follows. In section 2, first the SVM with explicit polynomial kernel is reviewed, then continuous output coding is presented and the proposed feature transformation as a combination of the two techniques is described. In section 3, the experimental setup and results are presented. Section 4 concludes the paper.

2. Discriminative transformation

We are interested in creating a set of new features for speech recognition, each element of the feature vector representing the discriminative information in distinguishing one phone from the others. In the proposed feature transformation system, the discriminative capabilities of SVM using an explicit sequence kernel will be combined with the continuous output coding (COC) technique [11]. By employing COC, we will be able to encode the information required to differentiate an individual phone from the rest in each dimension of the transformed feature space. In the following sections, we first describe the formulation of SVM with an explicit sequence kernel as a converter of input speech feature vectors to a scalar value measuring the similarity of the input frame to a phone and then, we illustrate the application of continuous output coding technique as the feature transform function.

2.1. SVMs using explicit polynomial expansion

SVM has shown to be effective in separating input vectors in 2-class problems [12], in which SVM effectively projects the vector $\mathbf{x}'$ into a scalar value $f(\mathbf{x}')$,

$$f(\mathbf{x}') = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}') + d \qquad (1)$$

where $y_i = \{-1, 1\}$, the vectors $\mathbf{x}_i$ are support vectors, N is the number of support vectors, α_i are adjustable weights, d is the bias term, and the function $K(\mathbf{x}_i, \mathbf{x}') = \phi(\mathbf{x}_i)^t . \phi(\mathbf{x}')$ is the kernel, in which $\phi(\mathbf{x})$ is a mapping from the input space to a

high dimensional space. If the explicit form of $\phi(\mathbf{x})$ is available, $f(\mathbf{x}')$ can be written as,

$$f(\mathbf{x}') = \left[\sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{y}_i) + \mathbf{d}\right]^t . \phi(\mathbf{x}') = \mathbf{w}_{svm}^t . \phi(\mathbf{x}') \quad (2)$$

where $\mathbf{d}=[d\ 0\ \ldots\ 0]$. In this new form, all the support vectors are collapsed down into a single model $\mathbf{w}_{svm}$ which is a weight vector in the high dimensional space. Directly applying SVMs on frame-level short-time spectral features involves high overlapping regions, resulting in a large number of support vectors. This problem can be alleviated by using an explicit kernel with the alternate form (2) for scoring.

In [13], an explicit kernel based upon comparing sequences of speech feature vectors for a measure of similarity, based on an expansion of short-time spectral feature space, has been adopted in speaker and language recognition. Having two sequences of short-time spectral features, $\mathbf{x}^M=\{\mathbf{x}_i | i=1,\ldots,M\}$ and $\mathbf{y}^L=\{\mathbf{y}_i | i=1,\ldots,L\}$, a kernel was constructed by training on one sequence of vectors using a generalized linear discriminant,

$$K(\mathbf{x}^M, \mathbf{y}^L) = \bar{\phi}_{\mathbf{x}}^t . \bar{\mathrm{R}}^{-1} . \bar{\phi}_{\mathbf{y}} \quad (3)$$

where the vector $\bar{\phi}_{\mathbf{x}} = \frac{1}{M}\sum_{i=1}^{M} \phi(\mathbf{x}_i)$ is the average vector over all expanded frames of the sequence $\mathbf{x}^M$ and $\bar{\mathrm{R}}$ is the correlation matrix derived from sequence vectors with different classes from target class of $\mathbf{x}^M$. Their system can be summarized as expanding the input feature vectors using polynomial expansion, and averaging sequences of high dimensional feature vectors and applying the explicit kernel, and finally performing classification using linear SVMs in the high dimensional space that showed successful results in speaker recognition application.

As for the feature transformation for speech recognition in this paper, we are interested in creating a set of new features, each element of the feature vector representing the discriminative information in distinguishing phones, rather than classification decisions. For each of the speech frames with a short-time spectral feature vector, we map them into a high dimensional space via polynomial expansion. Subsequently, we consider a window of consecutive frames of length M centered around the current high dimensional frame vector $\mathbf{x}$ to form sequence $\mathbf{x}^M$, and then $\bar{\phi}_{\mathbf{x}}$ is calculated over each sequence of consecutive frames $\mathbf{x}^M$. We aim at designing a SVM for distinguishing each of phones in speech recognition. Suppose $f_{tgt}(.)$ is output score of the SVM for the phone ***tgt***. From (2) and (3), with $\bar{\mathrm{R}} = \boldsymbol{I}$ for computational consideration, we have:

$$f_{tgt}(\mathbf{x}') = \left[\sum_{i=1}^{N} \alpha_i y_i \bar{\phi}_i^t + \mathbf{d}\right]^t . \bar{\phi}_{\mathbf{x}'} = \mathbf{w}_{tgt}{}^t . \bar{\phi}_{\mathbf{x}'} \quad (4)$$

2.2. Continuous output coding

It is not straightforward using SVMs in feature transformation for speech recognition, as $f(\mathbf{x}')$ is not a probability. We can use SVM output scores to approximate emission probabilities which can be adopted as input features for HMMs. In this way, we consider $f(\mathbf{x}')$ in (4) which can be interpreted as the similarity between the SVM hyperplane and the input vector $\mathbf{x}'$. The projection function is constructed by C SVM decision hyperplanes (e.g. linear discriminants), each of which trained to separate a specific phone from its competing phone set. In doing so, we denote the high dimensional vectors $\bar{\phi}_{\mathbf{x}}$ labeled with the target phone as the positive set, and the rest as the negative set. If we consider $\mathbf{w}_{\mathrm{svm}}$ for each trained SVM as a column of a $K\times C$ matrix $\mathbf{w}_{COC}$, where K is the dimension of the input vector $\bar{\phi}_{\mathbf{x}}$ and C is the number of SVMs, then the projection function can be written as,

$$\boldsymbol{\chi}' = \mathbf{w}_{COC}{}^t . \bar{\phi}_{\mathbf{x}'} \quad (5)$$

The C SVM outputs form a reduced dimensional space, which is also known as output coding [10]. The code size C of the so defined output code vector equals the number of phones. Output coding is a general method for solving multiclass problems by reducing them to multiple binary classification problems. It is able to correct some errors that individual classifiers make, thus also known as error-correcting output coding. Typically, output codes are defined as discrete codes of 0 and 1. Using SVM output as the output coding bit b, we have b=1 if $f(\mathbf{x}) > 0$, and b=0 otherwise. Some recent work improves the performance of output coding by relaxing the output codes from discrete coding to continuous coding [11]. Using the continuous coding means that points with low classification certainty contribute less to the score. In this paper, the continuous output coding (COC) projection function is implemented using (5).

After training a set of C independent SVM classifiers $\{f_1(.),f_2(.),...,f_C(.)\}$ while C is the number of phones, we project each high dimensional input vector $\bar{\phi}_{\mathbf{x}}$ to a vector of C real-valued SVM outputs $=\{f_1(\bar{\phi}_{\mathbf{x}}),f_2(\bar{\phi}_{\mathbf{x}}),...,f_C(\bar{\phi}_{\mathbf{x}})\}$ which is considered as the proposed discriminative features for subsequent phone recognition problem. Figure 1 shows how an input vector is exposed to C different SVM hyperplanes each trained to distinguish an individual phone. The corresponding output scores of SVMs are then gathered together to form the COC features. To have visual illustration, the simple case of 2-dimensional input vectors is considered in the Figure.

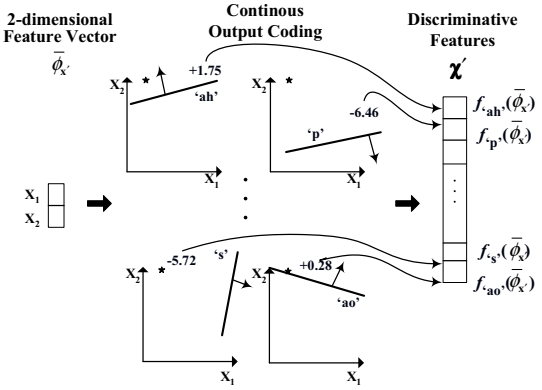

Figure 1. *Continuous output coding process*

Instead of original short-time spectral features, the resulting COC feature vectors are then fed into an HMM based speech recognition system. Figure 2 shows the framework of our proposed discriminative feature extraction system which is composed of three main stages. The two first steps namely, polynomial expansion and averaging in the high dimensional space constitute the explicit sequence kernel for each individual SVM. The third stage is the continuous output

coding to encode discriminative information from a set of C individual linear SVMs forming the proposed COC features for speech recognition.

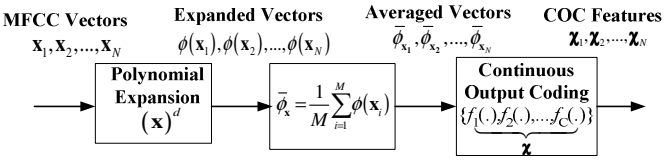

Figure 2. *Framework of discriminative feature extraction*

3. Experiments

In this section, we evaluate the discriminative capabilities of the proposed COC features. The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [14] and NTIMIT are used for the comparative study. NTIMIT is simply the TIMIT speech corpus which has been transmitted over a telephone network and re-recorded. The evaluations are conducted by the framewise phone classification with the high-quality manual phone labeling in TIMIT corpus, as well as the phone recognition task on TIMIT corpus. In order to assess the performance of COC features, we compare the results of phone classification and recognition using COC features to the results through the use of LDA features and MLP generated features. In the end, we investigate the robustness of the proposed COC features on NTIMIT corpus.

3.1. Experiment setup

The TIMIT corpus consists of 630 speakers, each speaker pronouncing 10 sentences. The 3969 **sx** and **si** sentences from the TIMIT proposed training set were used to train the phoneme models and the core test set consisted of 192 utterances from the standard 24-speaker was used for testing. We followed the common practice of HMM training for 48 phones and then mapping down to 39 phones for scoring purposes [15].

For the phone recognition system, decision-tree based state-tying context-dependent triphone models had been used for the acoustic modeling [16]. Approximately 1200 tied-states each having 16 mixture components have been built. A unigram phone language model was applied to the phone recognition. HTK toolkit [17] was used for both acoustic model training and phone recognition.

3.2. Discriminative feature vectors

For each speech frame, a 39-dimensional feature vector is extracted, consisting of 12 MFCCs and normalized energy, plus their first and second order derivatives. Sentence-based cepstral mean subtraction was applied to acoustic normalization both in train and test data.

To generate the proposed COC features, 39-dimensional MFCC vectors were expanded into high dimensional space by calculating all the monomials up to order 2, resulting $\binom{(39+1)+2-1}{2}=820$ elements of feature vector. It is expected that high dimensional feature vectors result in more discriminative capability to differ speech frames of one phone from others. Sequences of 9 successive frames, with 4 frames on each side of the current high dimensional frame, was used to obtain $\bar{\phi}_{\mathbf{x}}$. The obtained 820-dimensional feature vectors were used to train the 48 SVMs. Each SVM was trained in the one phone versus rest manner. For the SVM training, the high quality manually time-aligned phone labels of TIMIT corpus were used with the SVMTorch toolkit [18]. There are about 1.1 million labeled frames in the TIMIT training data set. In order to train the SVMs efficiently, the training vectors in each of the phone classes were quantized to 5000 centroids using k-means clustering algorithm. After training 48 independent SVM classifiers, each representative vector $\bar{\phi}_{\mathbf{x}}$ is mapped to a vector of 48 real-valued SVM outputs $\boldsymbol{\chi}=\{f_1(\bar{\phi}_{\mathbf{x}}), f_2(\bar{\phi}_{\mathbf{x}}),..., f_C(\bar{\phi}_{\mathbf{x}})\}$ as the COC features.

LDA features were acquired by LDA transformation on the nine-frame window of MFCC coefficients (9*39 dimensional vectors). MLP features are obtained by training an MLP network for discriminating phone classes. The number of the hidden layer nodes was set to 1000 and the number of the input nodes was 351 (9*39 dimensional input vectors). Softmax-activation function was used in the output layer during the training, but when using the MLP output as a feature vector to HMM, this nonlinearity was removed following the study in [7].

3.3. Framewise phone classification

To examine the discriminative capabilities of the COC features, we first conducted the framewise phone classification task while the classifications are made using one-versus-rest SVM classifiers. Figure 3 reports the experiments on different feature spaces, namely (i) the original MFCCs, (ii), LDA features (iii), MLP generated features, and (iv) the 48-dimension COC features. It is shown that the proposed discriminative features clearly outperform the MFCC, LDA, and MLP features due to more discrimination induced by the feature transformation process into the feature space.

By projecting the acoustic feature vectors into the 48 dimensions via the output codes of 48 SVMs with each SVM providing the discriminative information in the phone classification using explicit polynomial expansion, we successfully increased the discriminative capabilities at the acoustic feature level.

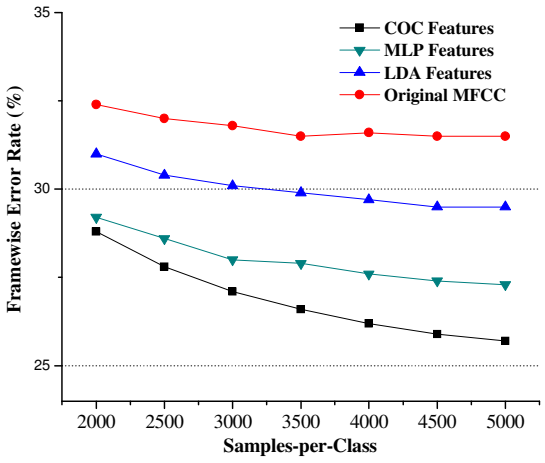

Figure 3: *Comparison of the framewise phone classification accuracy among different features.*

3.4. Phone recognition

In this section, the proposed COC features are fitted in the common adopted HMM framework. To measure the performance of COC features, we compare the results of the baseline phone recognizer on TIMIT task using COC features

to the results obtained through the use of original MFCCs, LDA features and MLP generated features. Note that in case of MLP feature transformation, we have also applied an additional stage of reducing the 48-dimensional MLP output vectors to 24 components by KLT (transformation matrix obtained from the training data) for a better performance [7]. Table 1 shows the Comparison of recognition error rates using different sets of features.

Table 1. *Comparison of recognition error rates using different generated feature vectors on TIMIT task.*

Phone Recognition	Phone Error
MFCC features	29.76 %
LDA features	29.35 %
MLP features	28.75 %
MLP + KLT features	28.13 %
COC features	26.88 %

The results show that COC features outperform other feature transformation techniques in phone recognition on TIMIT task. The improvement in recognition result reflects the effectiveness of using the proposed framework for the discriminative features in speech recognition systems.

3.5. Phone recognition with NTIMIT

In this section, we investigate the robustness of the proposed COC features on NTIMIT corpus. NTIMIT speech data have about 25dB SNR while TIMIT speech data have about 40dB SNR. In NTIMIT, there is also a great reduced spectral energy above 3.5 kHz due to telephone channel limitation. The recognition error rates using MFCC features and the COC features on NTIMIT corpus are compared in Table 2. It is shown that COC features outperform the MFCC features also in this telephone speech task.

Table 2. *Comparison of recognition error rates for NTIMIT.*

Phone Recognition	Phone Error
MFCC features	44.68 %
COC features	42.57 %

4. Conclusions

In this paper, a new approach for generating discriminative features for speech recognition was proposed. It was based on the fact that the commonly used short-time spectral features are not designed by optimizing a discriminative measure. We have introduced a new technique for feature transformation based upon explicit polynomial expansion. In order to benefit from the high-dimensional features, continuous output coding technique was applied to project the high-dimensional features to a much lower feature space, so that a conventionally-trained HMM based ASR system can be directly adopted. It has been shown that the discriminative capability of feature space has been improved by the proposed transformation process. Its application on both TIMIT and NTIMIT speech data showed that the COC feature transformation is effective in reducing recognition errors.

5. References

[1] Bahl, L. R., Brown, P. F., De Souza, P. V. and Merver R.L., "Maximum mutual information estimation of hidden Markov model parameters for speech recognition", Proc. ICASSP, 49-52, 1986.

[2] Juang, B. H., Chou, W. and Lee, C. H., "Minimum classification error rate methods for speech recognition", IEEE Trans. Speech and Audio Proc., 5(3):257-265, 1997.

[3] Povey, D. and Woodland, P.C., "Minimum phone error and I-smoothing for improve discriminative training", Proc. ICASSP, 105-108, 2002.

[4] Jiang, H., Li, X. and Liu, C., "Large margin hidden Markov models for speech recognition", IEEE Trans. Audio, Speech and Language Proc., 14(5):1584-1595, 2006.

[5] Haeb-Umbach, R. and Ney, H.,"Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition", Proc. ICASSP, I.13-I.16, 1992.

[6] Bourlard, H. and Morgan, N., "Connectionist Speech Recognition:A Hybrid Approach". Kluwer Academic Publishers, Boston, 1994.

[7] Hermansky, H., Ellis, D.P.W. and Sharma, S., "Tandem Connectionist Feature Extraction for Conventional HMM Systems", Proc. ICASSP, 1635-1638, 2000.

[8] Povey, D., Kingsbury, B., Mangu, L., Saon, G., Soltau, H., Zweig, G., "fMPE: Discriminatively Trained Features for Speech Recognition", Proc. ICASSP, 961-964, 2005.

[9] Droppo, J. and Acero., A., "Maximum Mutual Information SPLICE Transform for Seen and Unseen Conditions", Proc. interspeech, 989-992, 2005.

[10] Dietterich, T.G. and bakiri, G., "Solving multi-class learning problems via error-correcting output codes", Journal of Artificial Intelligence Research, 263-286, 1995.

[11] Crammer, K., and Singer, Y., "Improved Output Coding for Classification Using Continuous Relaxation", Proc. NIPS, 437-443, 2000.

[12] Vapnik, V., "The Nature of Statistical Learning Theory", Springer, N.Y., 1995.

[13] Campbell, W.M., Campbell, J.P., Reynolds, D.A., Singer E. and Torres-Carrasquillo, P.A., "Support vector machines for speaker and language recognition", Computer Speech and Language, 20(3):210–229, 2006.

[14] Fisher, W.M., Doddington, G.R. and Goudie-Marshall, K.M., "The DARPA speech recognition research database: specifications and status", Proc. DARPA Workshop on Speech Recognition, 93-99, 1986.

[15] Lee, K.F. and Hon, H.W., "Speaker-independent phone recognition using hidden Markov models", IEEE Trans. Acoust., Speech, Signal Processing, 37(11):1641-1648, 1989.

[16] Young, S., "Large vocabulary continuous speech recognition: review", IEEE Signal Processing Magazine, 13(5): 45-57, 1996.

[17] Young, S., Evermann, G., Kershaw, D., Odell, J., Ollason, D., Valtchev V. and Woodland, P., "The HTK Book (for HTK Version 3.2)", Cambridge University Engineering Dept, 2002.

[18] Collobert, R. and Bengio, S., "SVMTorch: Support vector machines for large-scale regression problems", Journal of Machine Lea rning Research, 1(2):143-160, 2001.