PHONEME RECOGNITION USING BOOSTED BINARY FEATURES

Anindya Roy[1,2], *Mathew Magimai.-Doss*[1], *Sébastien Marcel*[1]

[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

ABSTRACT

In this paper, we propose a novel parts-based binary-valued feature for ASR. This feature is extracted using boosted ensembles of simple threshold-based classifiers. Each such classifier looks at a specific pair of time-frequency bins located on the spectro-temporal plane. These features termed as Boosted Binary Features (BBF) are integrated into standard HMM-based system by using multilayer perceptron (MLP) and single layer perceptron (SLP). Preliminary studies on TIMIT phoneme recognition task show that BBF yields similar or better performance compared to MFCC (67.8% accuracy for BBF vs. 66.3% accuracy for MFCC) using MLP, while it yields significantly better performance than MFCC (62.8% accuracy for BBF vs. 45.9% for MFCC) using SLP. This demonstrates the potential of the proposed feature for speech recognition.

Index Terms— Phoneme recognition, automatic speech recognition, binary features, boosting.

1. INTRODUCTION

Phoneme/phone-specific information is embedded across both time and frequency in the speech signal. Standard ASR systems primarily use cepstral features which tend to capture the envelop of short-term magnitude spectrum of speech (frequency domain information). Dynamic information is subsequently added by appending approximate temporal derivatives of the cepstral features. Other features such as TRAPS/HATS [1], frequency domain linear prediction features [2], multiresolution RASTA features [3], 2D-DCT localized features [4] extract information directly from the spectro-temporal plane.

In this paper, we propose a novel parts-based approach for ASR which extracts binary (± 1) features from spectro-temporal segments of speech. This approach is inspired by Fern features which were successfully applied for object detection in computer vision [5]. For each phoneme, given equal-sized segments of spectro-temporal representation (in our case, log mel filter bank energies with temporal context of 170ms), the proposed approach builds simple binary classifiers for each time-frequency bin pair in the spectro-temporal segments. Then it selects (through boosting [6]) those bin pairs that best discriminates the phoneme against rest of the phonemes. Given a new spectro-temporal segment, the selected binary classifiers for each phoneme are applied on respective time-frequency bin pairs and their ± 1 decisions are used as input features for standard speech recognition system. We refer to these features as Boosted Binary Features (*BBF*).

The authors would like to thank the Swiss National Science Foundation, projects MultiModal Interaction and MultiMedia Data Mining (MULTI, 200020-122062) and Interactive Multimodal Information Management (IM2, 51NF40-111401) and the FP7 European MOBIO project (IST-214324) for their financial support.

The current work is partly motivated by a similar framework proposed by the authors for speaker verification (SV) task using only single frame information [7][8]. This framework yielded similar SV performance on clean condition and better performance on noisy conditions when compared to standard cepstral-based approach.

After feature extraction, there are two dominant approaches for acoustic modeling before integration into standard HMM-based ASR system, namely Gaussian mixture models (GMMs) and multilayer perceptrons (MLP). In this work, MLP was chosen as more suitable to model the binary-valued (± 1) inputs. We also tried single layer perceptrons (SLP) to model the binary features to verify our hypothesis that the phoneme classes could be linearly separable in this discriminative feature space.

We investigated the proposed feature on TIMIT phoneme recognition task. Our studies show that *BBF* yields performance similar or better than standard acoustic features using MLP. Using SLP, the proposed feature yields the least drop in performance and performs significantly better than standard features. The rest of the paper is organized as follows. In Sec.2, we describe the proposed binary features based framework. We describe our experiments in Sec.3. Finally, we discuss and outline the main conclusions of our work in Sec.4.

2. THE PROPOSED FRAMEWORK

2.1. Binary Features

In the first step, the input speech waveform is blocked into frames and processed via a bank of 24 Mel filters to yield a sequence of log spectral vectors of dimension $N_F = 24$. Sets of $N_T = 17$ consecutive such vectors are stacked to form spectro-temporal matrices of size $N_F \times N_T$.[1] Let $\mathbf{X}$ be such a spectro-temporal matrix. The (k,t)-th element, $X(k,t)$ of $\mathbf{X}$ denotes the log magnitude of k-th Mel filter output at t-th time frame. Consecutive spectro-temporal matrices are formed using shifts of one time frame, implying one spectro-temporal matrix per frame.

The proposed binary features are extracted from the matrix $\mathbf{X}$ as follows. A binary feature $\phi_i : \Re^{N_F \times N_T} \rightarrow \{-1, 1\}$ is defined completely by 5 parameters: two frequency indices, $k_{i,1}$, $k_{i,2} \in \{1, \cdots, N_F\}$, two time indices, $t_{i,1}$, $t_{i,2} \in \{1, \cdots, N_T\}$ and one threshold parameter, θ_i. The pairs of indices $(k_{i,1}, t_{i,1})$ and $(k_{i,2}, t_{i,2})$ define two time-frequency bins in the spectro-temporal matrix. To ensure two separate bins, both frequency and time indices should not be equal. The feature ϕ_i is defined as,

$$\phi_i(\mathbf{X}) = \begin{cases} 1 & \text{if } X(k_{i,1}, t_{i,1}) - X(k_{i,2}, t_{i,2}) \geq \theta_i, \\ -1 & \text{if } X(k_{i,1}, t_{i,1}) - X(k_{i,2}, t_{i,2}) < \theta_i. \end{cases} \quad (1)$$

In Fig. 1, we illustrate this process for an example 24×17 spectro-temporal matrix. Given the ranges of $k_{i,1}, k_{i,2}$ and $t_{i,1}, t_{i,2}$, the total

[1]In Sec. 3.3, we explain our choice of $N_T = 17$.

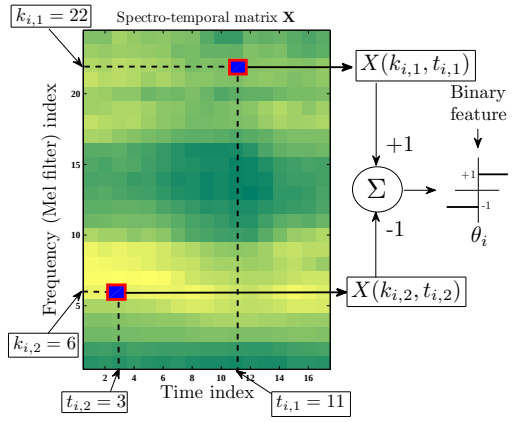

Fig. 1. Each binary feature ϕ_i is associated with a pair of time-frequency bins in the spectro-temporal matrix, defined by the parameters $(k_{i,1}, t_{i,1})$ and $(k_{i,2}, t_{i,2})$. The difference of the log magnitude values at these two bins is compared with a threshold θ_i and the sign is retained. An example feature ϕ_i is shown in the figure.

number of such binary features is $N_\Phi = N_T N_F (N_T N_F - 1)$. Let $\Phi = \{\phi_i\}_{i=1}^{N_\Phi}$ represent the complete set of such features.

2.2. Feature selection

Out of the complete set of binary features Φ, a certain number of features are iteratively selected *for each phoneme* according to their discriminative ability with respect to that phoneme. This selection is based on the Discrete Adaboost algorithm [6] with weighted sampling, which is widely used for such binary feature selection tasks [9] and is known for its robust performance [6]. These selected features are termed Boosted Binary Features (*BBF*). The boosting algorithm, which is to be run once for each phoneme, is as follows:

Algo.: Feature selection by Discrete Adaboost for a phoneme ω

Inputs: N_{tr} training samples, i.e. spectro-temporal matrices $\{\mathbf{X}_j\}_{j=1}^{N_{tr}}$ extracted from the training data; their corresponding class labels, $y_j \in \{-1, 1\}$, $(-1 : \mathbf{X}_j \notin \omega,\ 1 : \mathbf{X}_j \in \omega)$; N_f, the number of features to be selected; N_{tr}^*, the number of training samples to be randomly sampled at each iteration ($N_{tr}^* < N_{tr}$).

- Initialize the sample weights $\{w_{1,j}\} \leftarrow \frac{1}{N_{tr}}$.
- Repeat for $n = 1, 2, \cdots N_f$:
  - Normalize weights, $w_{n,j} \leftarrow \frac{w_{n,j}}{\sum_{j'=1}^{N_{tr}} w_{n,j'}}$
  - Randomly sample N_{tr}^* training samples, according to the distribution $\{w_{n,j}\}$
  - For each ϕ_i in Φ, choose threshold parameter θ_i to minimize misclassification error, $\epsilon_i = \frac{1}{N_{tr}^*} \sum_{j=1}^{N_{tr}^*} \mathbf{1}_{\{\phi_i(\mathbf{X}_j) \neq y_j\}}$ over the sampled set.
  - Select the next best feature, $\phi_n^* = \phi_{i^*}^*$ where $i^* = \arg\min_i \epsilon_i$
  - Set $\beta_n \leftarrow \frac{\epsilon_{i^*}}{1-\epsilon_{i^*}}$
  - Update the weights, $w_{n+1,j} \leftarrow w_{n,j} \beta_n^{\mathbf{1}_{\{\phi_n^*(\mathbf{X}_j) = y_j\}}}$

Output: The sequence of selected best features $\{\phi_n^*\}_{n=1}^{N_f}$.

Figure 2 illustrates the first 8 boosted features for phonemes /eh/, /ah/, /p/ and /s/, selected using training utterances from the TIMIT corpus. It can be observed that there are three distinct types of features:

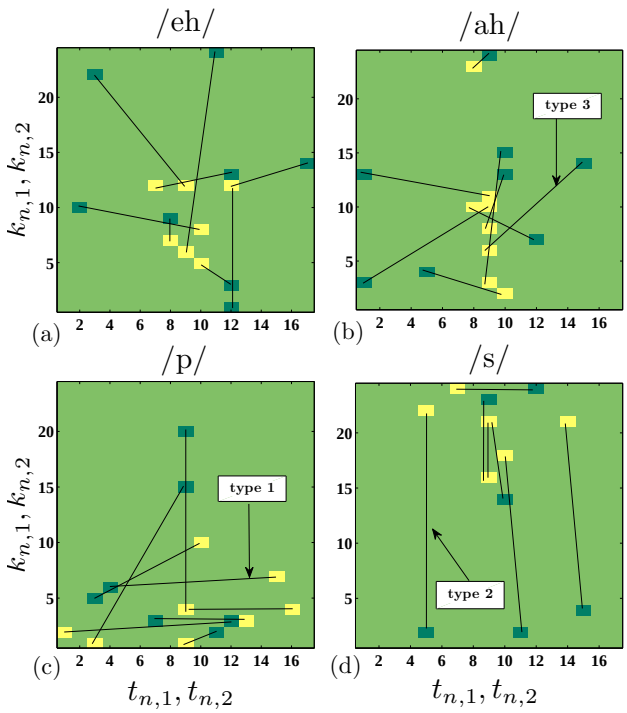

Fig. 2. Time-frequency bin pairs of the first 8 boosted features for phonemes /eh/, /ah/, /p/ and /s/ shown on the 24×17 spectro-temporal matrix. Horizontal axes denote time, vertical axes denote frequency, i.e. Mel filter indices. Each pair is indicated by a black line connecting the bin $(k_{n,1}, t_{n,1})$ (light yellow square) with the bin $(k_{n,2}, t_{n,2})$ (dark green square). One example of each of the 3 feature types are indicated. Please see Sec. 2.2 for details.

1. Features with time-frequency bins separated mostly in time. These features could be capturing similar temporal variation information as captured by TRAPS/HATS features in different frequency bands.
2. Features with bins separated mostly in frequency. These features could be capturing localized frequency information similar to cepstral features.
3. Features with bins separated along both time and frequency.

Hence, the proposed approach seems to present a general framework involving pairs of time-frequency bins on the spectro-temporal plane, some of which capture information along time, some along frequency and some along both, depending on their discriminative ability with respect to the particular phoneme being modelled.

For example, it is observed in Figure 2 that for fricative /s/ the features belong mostly to type 2 and are mainly in high frequency region, while for stop /p/ the features belong to type 1 and are mainly in low frequency region. For vowels, the features belong mostly to type 3, are closer to the center frame (in time) and lie mainly in the low to medium frequency region.

3. PHONEME RECOGNITION EXPERIMENTS

In this section, we describe the studies on TIMIT phoneme recognition task using our proposed framework.

3.1. Database

We use TIMIT acoustic-phonetic corpus for phoneme recognition experiments (excluding the SA sentences). The data consists of 3,000 training utterances from 375 speakers, 696 cross-validation utterances from 87 speakers, and 1,344 test utterances from 168 speakers. The 61 hand labeled phonetic symbols are mapped to set of 39 phonemes with an additional garbage class [10].

3.2. Features

We used a frame size of 25 ms and a frame shift of 10 ms to extract features. The features that are used in this study are:

1. *MFCC*: 39 dimensional acoustic feature vector consisting of 13 static Mel Frequency Cepstral Coefficients (MFCCs) with cepstral mean substraction and their approximate first order and second order derivatives (i.e., $c_0 - c_{12} + \Delta + \Delta\Delta$), extracted using HTK.
2. *MFBE*: 24 $\log$ **M**el **F**ilter **B**ank **E**nergies[2] over a context of 17 frames, i.e. a total of 408 features per frame. We study this feature as a holistic approach to compare with the proposed parts-based approach which involves spectro-temporal segments of the same size as *MFBE* but looks at only selected time-frequency bins (parts).
3. *BBF*: The proposed parts-based approach selects and trains binary features (termed **B**oosted **B**inary **F**eatures) from the spectro-temporal plane of $\log$ mel filter bank energies with temporal context of 17 frames (8 preceding and 8 following frames around the reference frame), i.e. a 24×17 spectro-temporal matrix (ref. Sec. 2.1).

   We used a subset of training data, more specifically 334 utterances (uniformly randomly chosen) out of the 3,000 utterances for selecting the binary features (described earlier in Sec. 2.2). This was done mainly to speedup the training process. The spectro-temporal matrices extracted from this data was split into two parts, namely, training samples and cross validation samples. The total number of training samples N_{tr} was 80,000 out of which the number of positive samples for each phoneme class was around 2,000. N^*_{tr}, the number of matrices randomly sampled at each boosting iteration was set to 4,000. The number of (selected) binary features N_f for each phoneme was set to 40 based on cross validation experiments (using 20,000 cross validation samples). This results in $40 \times 40 = 1600$ binary features per frame, aggregated over all phonemes.
4. *Rand*: To ascertain the utility of feature selection in our proposed parts-based approach, we also used features that involved randomly selected time-frequency bin pairs from the spectro-temporal plane. This was done in the following manner:
   (a) Create the complete set Φ of binary features considering all possible combinations of time-frequency pairs $(k_{i,1}, t_{i,1})$ and $(k_{i,2}, t_{i,2})$ (ref. Sec.2.2).
   (b) Uniformly randomly select 1600 features out of the set Φ.
   (c) For each of these 1600 binary features, compute the differences $X(k_{i,1}, t_{i,1}) - X(k_{i,2}, t_{i,2})$ over all training samples i.e. the same 80,000 samples used for selection and training of *BBF* feature. Simply set the median of these differences as the threshold θ_i for the feature.

   This results in a 1600-dimensional binary feature vector per frame, just as for the *BBF* features.

3.3. Classifier

We studied two different classifiers for each feature,

1. A single layer perceptron (SLP) classifier with softmax function for output units was trained to classify phonemes.
2. A multilayer perceptron (MLP) classifier was trained to classify phonemes in the conventional way.

In the case of *MFCC* feature, a 9 frame temporal context (4 frames of preceding and following context) was provided at the input of both SLP and MLP.

In the case of *MFBE* feature, a 17 frame temporal context (8 frames of preceding and following context) was provided at the input of both SLP and MLP. The choice of 17 frames is based on the total number of frames needed to estimate 9 frames of cepstral features with their first order and second order derivatives, where the derivative is estimated using 2 preceding and 2 following frames. This is also the reason why we restricted the spectro-temporal matrices to a temporal context of $N_T = 17$ in the case of *BBF*.

For *BBF*, the 1600-dimensional binary feature vector was provided at the input of both SLP and MLP.

The input dimension for each feature (for SLP and MLP) and number of hidden units (for MLP) is given in Table 1. In the case of *MFCC*, the number of hidden units was chosen based upon previous work reported in [11]. For *MFBE*, the hidden units were chosen so that the number of parameters are same as for *MFCC* feature based system. In the case of binary features, the hidden units were determined based on cross validation on the training data.

Feature	Input dimension	# of hidden units
MFCC	351	1000
MFBE	408	843
BBF	1600	400
Rand	1600	400

Table 1. Number of input units for SLP and MLP, and number of hidden units for MLP.

The SLPs and MLPs were trained using quicknet software[3]. The *MFCC* and *MFBE* features were normalized in the usual manner by global mean and standard deviation estimated on the training data. In the case of binary features, no normalization is done. The stopping criteria for training of SLP and MLP was frame accuracy on cross validation data of 696 utterances.

3.4. KL-HMM System

Conventional hybrid HMM/MLP based system use the output of MLP as local score (emission probabilities). In this work, we use Kullback Leibler (KL) divergence based acoustic modelling, where the probabilities of phoneme classes output of SLP and MLP are directly used as features. This system is referred to as KL-HMM system [12]. In KL-HMM, each state i is modeled by a multinomial distribution $\mathbf{y}_i = [y_i^1, \cdots, y_i^D]^T$, where D is the number of phonemes (in our case 40). Given a phoneme posterior feature observation (probabilities output by MLP or SLP) $\mathbf{z}_t = [z_t^1, \cdots, z_t^D]^T$

[2] from which the static MFCCs ($c_0 - c_{12}$) were extracted

[3] http://www.icsi.berkeley.edu/Speech/qn.html

at time t, the local score for state i is estimated as,

$$KL(\mathbf{y}_i, \mathbf{z}_t) = \sum_{d=1}^{D} y_i^d \log(\frac{y_i^d}{z_t^d})$$

The parameters of HMM (multinomial distributions) are trained using Viterbi expectation maximization algorithm with a cost function based on KL-divergence. The decoding is performed using standard Viterbi decoder. It has been shown that KL-HMM can perform better than hybrid HMM/MLP system.

In our studies, the KL-HMM was trained with the 3000 training utterances. Each phoneme was modeled by a three-state HMM. For recognition, the insertion penalties were tuned on cross validation data set, and then fixed for the test data.

3.5. Results

Table 2 presents the performance obtained for different features, in terms of phoneme recognition rate (obtained on the test data) and frame classification accuracy (obtained on the cross validation data).

Feature	SLP		MLP	
	CV Frame acc.	Phoneme rec. rate	CV Frame acc.	Phoneme rec. rate
MFCC	52.5	45.9	69.0	66.2
MFBE	52.4	46.6	68.2	66.6
BBF	**64.4**	**62.8**	**69.1**	**67.8**
Rand	59.5	56.2	67.3	65.0

Table 2. Frame accuracy on cross validation (CV) data and phoneme recognition rate on test set expressed in %.

The proposed *BBF* feature yields the best performance with both SLP and MLP. Interestingly, the *Rand* feature yields a close enough performance when compared to other features. It may be argued that the MLP system for *BBF* uses higher number of parameters than for *MFCC* and *MFBE* and hence yields better performance. So, in order to verify it, we trained MLPs for *MFCC* and *MFBE* features by increasing the number of hidden nodes to 1674 and 1462 respectively, to equalize the number of parameters. The performance for *MFCC* improved to 67.2% and for *MFBE* to 66.7%, which is still lower than the performance obtained with the proposed feature.

The study using SLP reveals interesting trends. The performance for *BBF* drops by 5% absolute (about 7.4% relative), whereas for *MFCC* and *MFBE*, it drops drastically i.e., 20.3% (about 30.6% relative) and 20.0% (about 30% relative) respectively. There is drop in performance for *Rand*, however, it is about 10% absolute better than *MFCC* and *MFBE*. Overall, these results support our initial hypothesis that the proposed binary features could be classified well using a linear classifier.

4. DISCUSSION AND CONCLUSIONS

It can be observed that the performance obtained with *MFCC* is lower than usually reported performance (of around 68%) in the literature [11][2] with hybrid HMM/MLP systems. This performance is achieved with speaker-level mean and variance normalization of the cepstral features. In this work, for fair comparison between features we did not perform speaker-level mean and variance normalization. However, the proposed binary feature approaches the performance reported in the literature. The reader may refer to [2] for comparison with more features.

The proposed *BBF* feature performs better than *Rand* thus showing the benefit of our boosting-based approach. However, *Rand* achieves acceptable performance, especially if the SLP performance is considered, where it performs significantly better than *MFCC* and *MFBE*. The extraction of both *BBF* and *Rand* in principle could be seen as a problem of finding a sparse representation for phoneme recognition. In the area of pattern recognition and signal processing, there are efforts towards finding such sparse representations. For example, in a recent work on face recognition, it has been shown that the choice of feature is less crucial if the sparsity of the recognition problem is harnessed properly [13]. Our studies may have implication towards this direction.

In this work, we used spectro-temporal representation derived from log mel filter bank energies. In principle, the extraction of *BBF* is not limited to spectro-temporal representation. For instance, it can be applied on phoneme posteriorgram (estimate of phoneme posterior probabilities across time). Also, we restricted our studies to a context of 17 frames for fair comparison with cepstral feature-based systems. The effect of using larger contexts for *BBF* could be investigated. Furthermore, we used equal number of binary features i.e. 40, for all phonemes. This may not be necessary. The number of binary features could possibly be decided for each phoneme in a data-driven manner. Future work will explore all these directions along with extension of our studies to conversational speech and speech corrupted by noise. The latter case could be specially interesting because such binary features have previously been shown to be robust against different types of noise for speaker verification task [7].

In summary, this preliminary work proposed a novel parts-based approach to extract binary features from the spectro-temporal plane. We evaluated the efficiency of the proposed feature on TIMIT phoneme recognition task. Our studies showed that the proposed binary features can yield performance similar or better than standard acoustic features.

5. REFERENCES

[1] S. Sharma and H. Hermansky, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. of ICASSP*, 1999.

[2] S. Ganapathy, S Thomas, and H Hermansky, "Static and dynamic modulation spectrum for speech recognition," in *Proc. of Interspeech*, 2009.

[3] H. Hermansky and P. Fousek, "Multi-Resolution RASTA Filtering for Tandem based ASR," *Proc. of Interspeech*, pp. 361–364, 2005.

[4] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized Spectro-Temporal Cepstral Analysis of Speech," in *Proc. of ICASSP*, 2008.

[5] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua, "Fast Keypoint Recognition using Random Ferns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, no. 3, pp. 448–461, 2010.

[6] J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," *Annals of Statistics*, vol. 28, pp. 2000, 1998.

[7] A. Roy, M. Magimai-Doss, and S. Marcel, "Boosted binary features for noise-robust speaker verification," in *Proc. of ICASSP*, 2010, pp. 4442–4445.

[8] A. Roy, M. Magimai-Doss, and S. Marcel, "A Parts-based Approach to Speaker Verification using Boosted Slice Classifiers," *Submitted to IEEE Trans. on Pattern Analysis and Machine Intelligence*, Sept. 2010.

[9] Y. Rodriguez, "Face Detection and Verification using Local Binary Patterns," PhD Thesis 3681, Ecole Polytechnique Federale de Lausanne, 2006.

[10] K-F Lee and H-W Hon, "Speaker-Independent Phone Recognition using Hidden Markov Models," *IEEE Trans. on Acoustics, Speech Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[11] J. Pinto, G.S.V.S Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP Based Hierarchical Phoneme Posterior Probability Estimator," *IEEE Transcations on Audio, Speech, and Language Processing*, 2010.

[12] G. Aradilla, H. Bourlard, and M. Magimai-Doss, "Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task," in *Proc. of Interspeech*, 2008.

[13] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Mar. 2008.