# BOOSTING LOCALIZED BINARY FEATURES FOR SPEECH RECOGNITION

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

[1]Spoken Language Processing group, LIMSI-CNRS, Orsay, France
[2]Idiap Research Institute, Martigny, Switzerland

## ABSTRACT

In a recent work, the framework of Boosted Binary Features (BBF) was proposed for ASR. In this framework, a small set of localized binary-valued features are selected using the Discrete Adaboost algorithm. These features are then integrated into a standard HMM-based system using either single layer perceptrons (SLP) or multilayer perceptrons (MLP). The features were found to perform significantly better (when coupled with SLP) and equally well (when coupled with MLP) compared to MFCC features on the TIMIT phoneme recognition task. The current work presents an overview of the idea and extends it in two directions: 1) fusion of BBF with MFCC and an analysis of their complementarity, 2) scalability of the proposed features from phoneme recognition to the continuous speech recognition task and reusability on unseen data.

***Index Terms***— Boosting, localized features, spectro-temporal features, speech recognition, feature fusion.

## 1. INTRODUCTION

Standard ASR systems primarily use cepstral features which tend to capture the envelop of short-term magnitude spectrum of speech. Dynamic information is subsequently added by appending approximate temporal derivatives of the cepstral features. These features are *holistic* (computed using the *whole* spectrum), *real-valued* and *based on prior knowledge* of the human speech production and perception systems.

In contrast, a novel set of *localized, binary-valued* ($\pm 1$) and *data-driven* features named Boosted Binary Features (BBF) was recently proposed for ASR [1][2], partly motivated by similar features proposed by the authors for the speaker verification task [3][4]. Each such feature is calculated by computing the difference in magnitude at two particular time-frequency bins in the spectro-temporal matrix ($\log$ mel filter bank energies with temporal context of 170ms): hence, the information is localized in time and frequency. This difference is compared with a threshold and the result of this comparison gives a binary value which is taken as the feature. Considering *all* possible pairs of time-frequency bins, a large set of binary features is created. The Adaboost algorithm [5] selects a small subset of these features which best discriminate a particular phoneme against all other phonemes, based on some training data. These selected features are termed as Boosted Binary Features (BBF). After extraction, the features are modelled by multilayer perceptrons (MLP) or single layer perceptrons (SLP)[1]. The posterior estimates of MLP and SLP are then provided to a Kullback-Leibler Hidden Markov Model for decoding [6].

Previously, phoneme recognition studies on the TIMIT database [1] showed that BBF achieved a phoneme recognition rate of $67.8\%$ which is slightly better than $66.2\%$ obtained by cepstral features under similar conditions, using MLP. Using SLP, BBF performed significantly better at $62.8\%$ than cepstral features at $45.9\%$.

In the current work, we extend these studies in the following directions:[1] 1) Analysis of the possible complementary nature of BBF and cepstral features, and fusion of the two at a) feature level and b) decision level. 2) Extension from phoneme recognition to continuous speech recognition task (i.e. word recognition), and testing the generalization capability of BBF to unseen data.

The rest of the paper is organized as follows. In Sec.2, we describe the BBF framework and its integration via SLP and MLP to a HMM system. We describe the fusion experiments in Sec.3 and the continuous speech recognition studies in Sec.4. Finally, we outline the main conclusions of the work and discuss future directions in Sec.5.

## 2. BRIEF THEORY

### 2.1. Binary features and their selection

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 $N_F \times N_T$.[2] Let $\mathbf{X}$

The first author completed the work while at the Idiap Research Institute, Martigny, Switzerland. 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 FP7 European MOBIO project (IST-214324) for their financial support.

[1] Note that the above studies were first reported in [2]. We summarize the studies here for a wider dissemination of this novel work.

[2] $N_T = 17$ is chosen to ensure a fair comparison with MFCC features: a temporal context of 17 frames is required to compute 9 frames of (MFCC +

be such a spectro-temporal matrix. The $(k,t)$-th element, $X(k,t)$ of $\mathbf{X}$ denotes the $\log$ magnitude of the $k$-th Mel filter output at $t$-th time frame. The proposed binary features are extracted from the matrix $\mathbf{X}$ as follows. A binary feature $\phi_i : \Re^{N_F \times N_T} \to \{-1,1\}$ 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, $\theta_i$. 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. The feature $\phi_i$ is defined as,

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

Given the ranges of $k_{i,1}, k_{i,2}$ and $t_{i,1}, t_{i,2}$, the total number of such binary features is quite huge: $N_\Phi = N_T N_F (N_T N_F - 1) = 17 \cdot 24 \cdot (17 \cdot 24 - 1) \approx 1.7 \times 10^5$. Out of all these features, a certain number $N_f$ ($\approx 40$) are selected *for each phoneme* according to their discriminative ability with respect to that phoneme (1-*vs.*-all classification) given a set of training examples. This selection is based on the Discrete Adaboost algorithm with weighted resampling [5]. Details about the algorithm are provided in [1][2]. Features selected for all phonemes are aggregated and termed as Boosted Binary Features (BBF). This forms a vector $\mathbf{f}$ of binary values of dimension $D = N_f \times N_\Omega$ where $N_\Omega$ is the number of phonemes used.

### 2.2. Phoneme posterior probability estimation

In this work, single layer perceptrons (SLP) and multilayer perceptrons with one hidden layer (MLP) are used as posterior feature estimators. The input to the SLP or MLP is the BBF vector $\mathbf{f}$ described before. Outputs are the posterior probability estimates for the phonemes, $\mathbf{z}_t = [z_t^1, \cdots, z_t^{N_\Omega}]^T$ at every time step $t$.

### 2.3. KL-HMM system

The phoneme posterior probability estimates of SLP and MLP are used as feature observations in a Kullback Leibler divergence-based Hidden Markov Model (KL-HMM) system [6]. [3] In KL-HMM, each state $i$ is modeled by a multinomial distribution $\mathbf{y}_i = [y_i^1, \cdots, y_i^{N_\Omega}]^T$, where $N_\Omega$ is the number of phonemes. Given a phoneme posterior feature observation $\mathbf{z}_t = [z_t^1, \cdots, z_t^{N_\Omega}]^T$ at time $t$, the local score for state $i$ is estimated as the Kullback Leibler divergence between $y_i$ and $z_t$, i.e.,

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

The parameters of KL-HMM (multinomial distributions) are trained using Viterbi EM algorithm with a cost function based on KL-divergence. Each phoneme is modeled by a three-state HMM.

## 3. FUSION STUDIES: BBF AND MFCC

Due to the contrasting properties of BBF (localized and binary) and standard cepstral features (holistic and real-valued), they could carry useful complementary information, possibly suitable for fusion.

### 3.1. Analysis of complementary nature

We analysed the possible complementary nature of BBF and MFCC in two ways. In the first approach, we trained two MLPs, one with BBF and the other with MFCC as inputs, using 3000 training utterances from the TIMIT database [7]. In Table 1, we show the distribution of frames[4] from the 696 cross-validation utterances from TIMIT according to whether they were correctly or incorrectly classified by these two MLPs into one of the $N_\Omega = 40$ phonemes in TIMIT. To classify, the phoneme with the maximum MLP posterior estimate was selected. We observe that 8.8% of the frames were incorrectly classified with MFCC but correctly classified with BBF, and 9.2% of the frames were incorrectly classified with BBF but correctly classified with MFCC. This shows that BBF could rectify some errors of MFCC, and vice-versa. This indirectly suggests that BBF and MFCC carries useful complementary information.

In the second approach, we consider a representative subset of the 40 phonemes in the TIMIT database and analyse the frame-level phoneme classification accuracy of BBF and MFCC for each (ref. Table 2). It is observed that MFCC performs better than BBF for vowels /ay/ and /ih/, liquid /l/ and nasal /m/ while BBF outperforms MFCC for fricatives /th/, /hh/,/v/ and /f/. Again this indirectly suggests the complementarity of the two features in the sense that one seems to carry more discriminative information related to certain phoneme types while the other carries more discriminative information related to other phoneme types.

### 3.2. Fusion experiments

Two systems based on the fusion of these two features were studied:

1. **Feature-level fusion:** A 1600-dimensional BBF feature vector [5] is concatenated with 351-dimensional MFCC feature vector (i.e. 39-dimensional MFCC+$\Delta$+$\Delta\Delta$ vectors accumulated over a context of 9 frames) to form a 1951-dimensional fused feature vector. This is modeled by an MLP and the posterior probabilities estimated by the MLP are used as observations in a KL-HMM system.

$\Delta$MFCC + $\Delta\Delta$MFCC) [1].

[3] The KL-HMM was chosen because it matches well with posterior estimates of MLP/SLP and achieves performance better than hybrid HMM/MLP systems and comparable to standard HMM/GMM systems [6].

[4] extracted using standard frame size of 25ms and frame shift of 10ms.

[5] ref. Section 2.1, $D = N_f \times N_\Omega$, $N_f = 40$, $N_\Omega = 40$.

|  | BBF correct | BBF incorrect |
|---|---|---|
| MFCC correct | 61.5 | 9.2 |
| MFCC incorrect | 8.8 | 20.5 |

**Table 1**. Distribution (%) of frames from cross-validation set of TIMIT database.

2. **Decision-level fusion:** Two MLPs were trained individually using only BBF and only MFCC features. Their phoneme posterior probability estimates were then dynamically combined via Dempster-Shafer method described in [8]. Subsequent modeling via KL-HMM was the same as before.
These two systems were evaluated on the phoneme recognition task using the TIMIT database. Note that in this case, the number of hidden units of the MLP in each system was set so that the total number of parameters was constant over *all* the systems, in order to ensure a fair comparison.Table 3 shows the performance obtained for different systems (fusion and individual) in terms of phoneme recognition rate (PRR) obtained on the 1344 test utterances of TIMIT and frame classification accuracy obtained on the cross-validation utterances of TIMIT.

The following points related to the results reported are noteworthy: 1) The fusion of MFCC with BBF is beneficial. It leads to a 3% increase in PRR over MFCC and a 1.1% increase over BBF individually. 2) Both decision fusion *and* feature fusion perform better than the individual feature-based systems. These observations support the hypothesis that BBF and MFCC could contain useful complementary information and a combination of these two features results in improved ASR performance.

## 4. CONTINUOUS SPEECH RECOGNITION STUDIES

This section investigates: a) the scalability of these features from the phoneme recognition task [1] to the continuous speech recognition task (i.e. word recognition), and b) the use of auxiliary data to select the features.

| Phoneme | Accuracy (%) |  | Best | Improvement (%) |  |
|---|---|---|---|---|---|
|  | MFCC | BBF | feature | Absolute | Relative |
| /ay/ | 71.8 | 64.3 | MFCC | 7.5 | 10.4 |
| /ih/ | 68.4 | 61.9 | MFCC | 6.4 | 9.4 |
| /l/ | 70.5 | 66.0 | MFCC | 4.5 | 6.4 |
| /m/ | 66.9 | 63.2 | MFCC | 3.6 | 5.5 |
| /th/ | 24.5 | 31.6 | BBF | 7.1 | 22.4 |
| /hh/ | 59.7 | 66.5 | BBF | 6.8 | 10.2 |
| /v/ | 54.0 | 60.0 | BBF | 6.0 | 10.1 |
| /f/ | 78.6 | 82.7 | BBF | 4.1 | 5.0 |

**Table 2**. Best feature and relative improvement in frame accuracy on cross-validation set of TIMIT database for subset of phonemes.

| System | CV Frame Accuracy | Phoneme Rec. Rate (PRR) |
|---|---|---|
| BBF only | 70.3 | 69.3 |
| MFCC only | 69.9 | 67.4 |
| Feature fusion | 70.6 | **70.4** |
| Decision fusion | 73.2 | **70.3** |

**Table 3**. Results of different systems using MFCC, BBF and fusion of the two (in %) on TIMIT.

### 4.1. Database and experiment setup

The DARPA Resource Management (RM) corpus [9] was used. It consists of read queries on the status of naval resources. The corpus is partitioned into training set (2,880 utterances), development set (1,110 utterances) and evaluation set (1,200 utterances) [10] and has a vocabulary of 991 words. The phoneme-based lexicon was obtained from the UNISYN dictionary. There are 45 context-independent phonemes including silence. A frame size of 25 ms and a frame shift of 10 ms was used to extract features. The features used in this study are: 1) **MF-PLP**: 39 dimensional feature vector consisting of 13 static Mel Frequency PLP Cepstral Coefficients (MF-PLP) with cepstral mean subtraction and their approximate first and second order derivatives (i.e., $c_0 - c_{12} + \Delta + \Delta\Delta$). 2) **BBF**: Two sets of BBF were considered, as follows. a) BBF-TIMIT The first 80,000 samples (spectro-temporal matrices) extracted from training partition of TIMIT database is used as training data to select the features (ref. Section 2.1). The purpose is to evaluate the generalization capability of these features boosted using TIMIT to a speech recognition task using a different database, RM. b) BBF-RM In a similar way, the first 80,000 samples extracted from the training partition of the RM database is used to select the features. In this case, the feature selection and speech recognition studies use the *same* database. 3) **Rand**: To ascertain the utility of the feature selection algorithm, we also used features that involved *randomly selected* time-frequency bin pairs from the spectro-temporal plane[1]. As with BBF, two cases are considered: a) Rand-TIMIT, and b) Rand-RM. As before, the phoneme posterior estimates are sent as feature observations to a KL-HMM system. Two types of KL-HMM systems are considered: 1) context-independent sub-word unit based system, and 2) word internal context-dependent sub-word unit based system [10].

### 4.2. Results

The performance obtained for different features in terms of word error rate (WER) on the evaluation set of the RM corpus is reported in Table 4, for context-independent and context-

| Feature | Context independent | | Context dependent | |
|---|---|---|---|---|
| | MLP | SLP | MLP | SLP |
| MF-PLP | 7.1 | 28.3 | 5.1 | 14.7 |
| BBF-TIMIT | 7.6 | 11.1 | 5.5 | 7.1 |
| BBF-RM | 7.8 | 10.9 | 5.6 | 7.2 |
| Rand-TIMIT | 9.2 | 17.5 | 6.8 | 10.3 |
| Rand-RM | 9.2 | 16.8 | 6.4 | 10.8 |

**Table 4**. Word Error Rate (%) on evaluation set of RM database.

dependent systems. The following points are noteworthy: 1) In general, context-dependent systems show a reduction in WER over context-independent systems. 2) With MLP, BBF and MF-PLP perform comparably well, with WERs ranging from 5.1 to 5.6% for context-dependent, and 7.1 to 7.8% for context-independent. As reported in [10], standard HMM/Gaussian Mixture Model system and Tandem features based system (which are equivalent in terms of context modeling to the context-dependent system reported here) achieve 5.7% WER each. This is similar to the WER achieved using BBF. 3) BBF-TIMIT and BBF-RM show similar performance. This shows that BBF is not sensitive to the training data used for boosting, and can generalize well to unseen data. 4) Going from MLP to SLP, BBF shows significantly lower degradation in performance compared to MF-PLP in all cases. For example, WER for BBF-TIMIT increases from 5.5 to 7.1 %, i.e. a relative increase of 29 %, while WER for MF-PLP increases from 5.1 to 14.7 %, a relative increase of 188 %, for the context-dependent case. 5) Rand features also achieve reasonable performance. Interestingly, in case of SLP they perform better than MF-PLP. However, they perform worse than BBF in *all* cases, showing the utility of the feature selection stage.

## 5. CONCLUSIONS AND FUTURE WORK

Firstly, the fusion of Boosted Binary Features with cepstral features led to an improvement in ASR performance at both the feature level and the decision level, possibly drawing from the complementary nature of the two feature types. Secondly, continuous speech recognition experiments using the Resource Management database showed that BBF compares well with cepstral features in terms of word error rate. Importantly, although the framework is data-driven, our study suggests that the performance of BBF is independent of the dataset used to select the features.

From a machine learning perspective, possible directions for future work are as follows: 1) In this work, a one-vs-all strategy was used to select the binary features. Other selection strategies could be explored, such as feature sharing across classes and multiclass boosting strategies [11]. 2) Instead of depending on MLPs, phoneme posterior inputs to KL-HMM could be directly modeled by extending the system from BBF to boosted decision trees. 3) The extraction of BBF could be interpreted as adding a layer to the MLP or SLP to learn phone-specific representations directly from spectro-temporal plane using auxiliary data. This could complement deep-learning frameworks geared towards similar objectives [12]. Other general directions are as follows: 1) When previously applied to speaker recognition, BBF showed better noise-robustness than cepstral features [4]. It is worthwhile to verify this property of BBF for ASR also. 2) Fusion studies should be extended from phoneme recognition to a complete word recognition task.

## 6. REFERENCES

[1] A. Roy, M. Magimai-Doss, and S. Marcel, “Phoneme Recognition using Boosted Binary Features,” in *Proceedings of ICASSP*, 2011, pp. 4868–4871.

[2] A. Roy, *Boosting Localized Features for Speaker and Speech Recognition*, Ph.D. thesis, Ecole Polytechnique Federale de Lausanne, Nov. 2011.

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

[4] A. Roy, M. Magimai.-Doss, and S. Marcel, “A Fast Parts-based Approach to Speaker Verification using Boosted Slice Classifiers,” *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 241–254, 2012.

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

[6] G. Aradilla, H. Bourlard, and M. Magimai-Doss, “Using KL-Based Acoustic Models in a Large Vocabulary Recognition Task,” in *Proceedings of Interspeech*, 2008, pp. 928–931.

[7] W.M Fisher, G.R. Doddington, and K.M. Goudie-Marshall, “The DARPA speech recognition research database: Specifications and status,” *Proceedings of DARPA Workshop on Speech Recognition*, pp. 93–99, Feb. 1986.

[8] F. Valente, “A Novel Criterion for Classifiers Combination in Multistream Speech Recognition,” *IEEE Signal Processing Letters*, vol. 16, no. 7, pp. 561–564, 2009.

[9] P. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett, “The DARPA 1000-word resource management database for continuous speech recognition,” in *Proceedings of ICASSP*, 1988, pp. 651–654.

[10] J. Dines and M. Magimai.-Doss, “A study of phoneme and grapheme based context-dependent ASR systems,” in *Proceedings of MLMI 2007, Lecture Notes in Computer Science, 4892*, 2008, pp. 215–226.

[11] J. Zhu, H. Zou, S. Rosset, and T. Hastie, “Multi-class AdaBoost,” *Statistics and Its Interface*, vol. 2, pp. 349–360, 2009.

[12] A. Mohamed, G. Dahl, and G. E. Hinton, “Acoustic modeling using deep belief network,” *IEEE Transactions on Audio, Speech, and Language Processing (in press)*, 2012.