IDIAP RESEARCH REPORT


# CONTINUOUS SPEECH RECOGNITION USING BOOSTED BINARY FEATURES

Anindya Roy  Mathew Magimai.-Doss
Sébastien Marcel

Idiap-RR-35-2011

OCTOBER 2011

**Centre du Parc**, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11  F +41 27 721 77 12  info@idiap.ch  www.idiap.ch

# Continuous Speech Recognition using Boosted Binary Features

Anindya Roy*, *Student Member, IEEE,* Mathew Magimai.-Doss, *Member, IEEE,* and Sébastien Marcel, *Member, IEEE*

***Abstract*—A novel parts-based binary-valued feature termed Boosted Binary Feature (BBF) was recently proposed for ASR. Such features look at specific pairs of time-frequency bins in the spectro-temporal plane. The most discriminative of these features are selected by boosting and integrated into a standard HMM-based system using multilayer perceptron (MLP) and single layer perceptron (SLP). Previous studies on TIMIT phoneme recognition task showed that BBF yields similar or better performance compared to cepstral features. In this work, this study is extended to continuous speech recognition task on the DARPA Resource Management database. Results show that BBF achieves comparable word error rate (5.5%) on this task with respect to standard cepstral features (5.1%) using MLP. Using SLP, the error rate for BBF shows much lower degradation (from 5.5% to 7.1%) compared to cepstral features (from 5.1% to 14.7%). In addition, it is found that BBF features can be selected well using auxiliary data.**

***Index Terms*—EDICS: SPE-RECO, SAS-MALN, SAS-STAT**

## I. Introduction

STANDARD automatic speech recognition (ASR) systems use different types of features such as cepstral features and their approximate temporal derivatives, TRAPS/HATS [1], multiresolution RASTA features [2] and 2D-DCT localized features [3]. These features capture in various ways phoneme-specific information embedded across time and frequency.

In this context, a set of parts-based binary $(\pm 1)$ features was recently proposed for ASR [4] which present a general framework to capture phoneme-specific information embedded 1) across frequency, 2) across time, and 3) across both time and frequency. These features are related to but distinct from local features used for isolated digit recognition in [5] and are inspired by similar binary features which have been successfully applied for face and object detection in the computer vision domain [6][7].

These parts-based features are extracted by computing the difference in magnitude at two time-frequency bins (i.e. the *parts*) in a spectro-temporal matrix formed by stacking $\log$ mel filter bank energies over a temporal context of 170ms, and comparing this difference with a threshold. The binary $(\pm 1)$ result of this comparison is taken as the feature. Considering *all* possible pairs of time-frequency bins and *all* possible thresholds, a very large set of binary features is created. Out of this set, the Adaboost algorithm [8] is used to select a small number of features which best discriminate a particular phoneme against all others. These selected features, termed as **B**oosted **B**inary **F**eatures (*BBF*), are modelled by multilayer perceptron (MLP) or single layer perceptron (SLP)[4]. The phoneme posterior probabilities estimated by MLP or SLP are then used as feature observation for Kullback Leibler divergence based Hidden Markov Model (KL-HMM) system [9].

Previous studies on TIMIT phoneme recognition task showed that *BBF* yields performance similar or better than standard cepstral features [4]. This paper investigates: a) the scalability of these features to continuous speech recognition task, and b) use of auxiliary data to select the features. On DARPA Resource Management (RM) database, our studies show that: a) *BBF* can yield performance comparable to standard cepstral features using MLP, b) using SLP, *BBF* performance degrades significantly *less* compared to cepstral features, supporting the observation made in previous studies [4], and c) *BBF* features selected using an auxiliary corpus can yield same performance as those selected using the task specific data.

The rest of the paper is organized as follows. In Sec.II, we describe the *BBF* based framework. We describe the experimental setup in Sec.III. The results from our speech recognition studies are detailed in Sec.IV. Finally, we discuss and outline the main conclusions of the work in Sec.V.

## II. Brief Theory

The theory of boosted binary features for speech recognition was proposed in a previous work [4]. Since it is relatively new, it is described again for convenience.

### A. 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 the $k$-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 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 by 5 parameters: two frequency indices, $k_{i,1},\ k_{i,2} \in \{1, \cdots, N_F\}$, two time indices,

A. Roy is with the École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, and Idiap Research Institute, Martigny, Switzerland. E-mail: Anindya.Roy@idiap.ch

M. Magimai.-Doss and S. Marcel are with Idiap Research Institute, Martigny, Switzerland.

[1] In Sec. III-C, the reason behind the choice of $N_T = 17$ is explained.

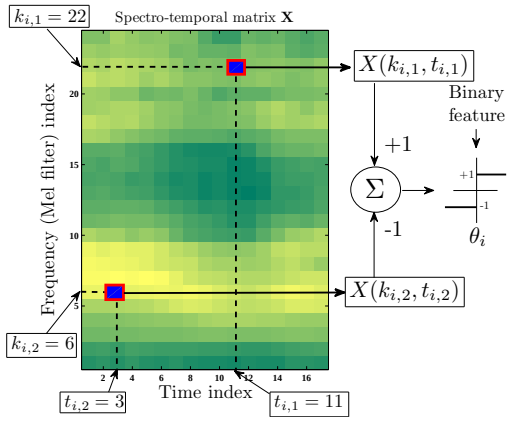


Fig. 1. Each binary feature $\phi_i$ is associated with a pair of time-frequency bins in the spectro-temporal matrix, defined by the parameters $(k_{i,1}, t_{i,1})$ and $(k_{i,2}, t_{i,2})$. The difference of the log magnitude values at these two bins is compared with a threshold $\theta_i$ and the sign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

$t_{i,1},\ t_{i,2} \in \{1, \cdots, N_T\}$ 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. To ensure two separate bins, both frequency and time indices should not be equal. The feature $\phi_i$ is defined as,

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

In Fig. 1, we illustrate this process for an example $24 \times 17$ spectro-temporal matrix. 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 $N_\Phi = N_T N_F (N_T N_F - 1) = 17 \cdot 24 \cdot (17 \cdot 24 - 1) \approx 1.7 \times 10^5$. Let $\Phi = \{\phi_i\}_{i=1}^{N_\Phi}$ denote the complete set of such features.

### B. Binary Feature Selection

Out of the complete set of binary features $\Phi$, a certain number of features $N_f$ ($\approx$ 40) are selected *for each phoneme* according to their discriminative ability with respect to that phoneme, using Discrete Adaboost algorithm [8] with weighted resampling, which is widely used for such feature selection tasks [10] and is known for its robust performance [8]. The process is data-driven and requires training data.[2] The boosting algorithm, which is to be run once for each phoneme, is described next.

**Feature selection algorithm by Discrete Adaboost for a phoneme** $\omega$

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 $(N_{tr}^* < N_{tr})$.[3]

- 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 $\phi_i$ in $\Phi$, choose threshold parameter $\theta_i$ 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 $\{\phi_n^*\}_{n=1}^{N_f}$.

After the selection process, the features selected for all phonemes are aggregated and termed as **B**oosted **B**inary **F**eatures (*BBF*). This forms a vector of binary values of dimension $D = N_f \times N_\Omega$ where $N_\Omega$ is the number of phonemes considered. The reader may refer to [4] for an analysis of the features selected by Adaboost for different phonemes.

## III. Experimental Setup

In this section, we describe the setup for our continuous speech recognition experiments using *BBF* and cepstral features.

### A. Database

The DARPA Resource Management (RM) corpus [11] is used for the experiments. The RM corpus consists of read queries on the status of naval resources. The database is partitioned into training set (2,880 utterances), development set (1,110 utterances) and evaluation set (1,200 utterances) [12]. Training and development utterances are spoken by 109 speakers and correspond to approximately 3.8 hours of speech data. Evaluation set amounts to 1.1 hours of speech data and is covered by a word pair grammar included in the task specification. RM corpus has 991 words. The phoneme-based lexicon was obtained from the UNISYN dictionary. There are 45 context-independent phonemes including silence.

### B. 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) *MF-PLP*: 39 dimensional feature vector consisting of 13 static Mel Frequency PLP Cepstral Coefficients (MF-PLP) with cepstral mean substraction and their approximate first and second order derivatives (i.e., $c_0 - c_{12} + \Delta + \Delta\Delta$), extracted using HTK.
2) *BBF*: **B**oosted **B**inary **F**eatures are extracted from spectro-temporal matrices of size $24 \times 17$ (ref. Sec. II). Two sets of *BBF* are considered:
   a) *BBF-TIMIT* The first 80,000 samples (spectro-temporal matrices) extracted from training partition of TIMIT database [13] is used as training data to select the features (ref. Sec.II-B).[4] The

[2] In Sec. IV, it is shown that the choice of training data is not important and the selected features can generalize well to unseen data.

[3] Values of $N_{tr}^* \approx 5\%$ of $N_{tr}$ works sufficiently well. It was not tuned.

[4] Using a subset rather than *all* samples ($\approx 1.4 \times 10^6$) led to faster boosting with no loss in performance.

purpose is to evaluate the generalization capability of these features boosted using TIMIT [4] to a speech recognition task using a different database, RM. The TIMIT data is labeled using $N_\Omega = 40$ phoneme classes. $N_f = 40$ binary features are selected for each phoneme [4], leading to a feature vector of dimension $D = N_f \times N_\Omega = 40 \times 40 = 1600$ per frame.

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 these features. In this case, the feature selection and speech recognition studies use the *same* database. The RM data is labeled using $N_\Omega = 45$ UNISYN phoneme classes, leading to a feature vector of dimension $D = 40 \times 45 = 1800$ per frame.

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. This was done in the following manner [4]: a) Create the complete set $\Phi$ 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.II-A). b) Uniformly randomly select required number of features out of the set $\Phi$. c) For each of these 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 of *BBF* feature. Simply set the median of these differences as the threshold $\theta_i$ for the feature.

As for *BBF*, two cases are considered: a) *Rand-TIMIT* The training samples were extracted from the TIMIT database. b) *Rand-RM* The training samples were extracted from the RM database.

We compare these features by first training a phoneme class conditional probability estimator using these features as input, and using the estimates of phoneme a posteriori probabilities, referred to as *posterior features*, as feature observations for KL-HMM system.

### C. Posterior Feature Estimation

Similar to our previous work [4], we studied two different posterior feature estimators for each acoustic 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 *MF-PLP* feature, a 9 frame temporal context (4 frames of preceding and following context), i.e. a 9×39 = 351-dimensional feature vector was provided at the input of SLP and MLP. This explains the choice of 17 frames for *BBF*. It is to ensure a fair comparision, 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.

In the case of *MF-PLP*, an off-the-shelf MLP trained on exactly the same setup was used [12]. For *BBF* and *Rand*, the 1600 or 1800-dimensional *BBF* vector was provided at the input of both SLP and MLP. The number of hidden units for MLP was determined by cross-validation based on frame-level phoneme accuracy obtained on RM development set. The SLPs and MLPs were trained using quicknet software[5]. The *MF-PLP* 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 criterion for training of SLP and MLP was frame-level phoneme accuracy on the development set. Table I shows the frame-level phoneme accuracy obtained for different features on the development set.

| Feature | MLP | SLP |
|---|---|---|
| *MF-PLP* | 73.2 | 54.2 |
| *BBF-TIMIT* | 73.1 | 65.6 |
| *BBF-RM* | 72.8 | 65.9 |
| *Rand-TIMIT* | 70.9 | 59.3 |
| *Rand-RM* | 71.0 | 60.3 |

TABLE I
FRAME-LEVEL PHONEME ACCURACY (%) ON RM DEVELOPMENT SET.

### D. KL-HMM System

The posterior features estimated by MLP or SLP are used as feature observations in the framework of KL-HMM system [9]. Briefly, 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 (in this case 45). Given a phoneme posterior feature observation (e.g., probabilities estimated by MLP), $\mathbf{z}_t = [z_t^1, \cdots, z_t^{N_\Omega}]^T$ at time $t$, the local score for state $i$ is estimated as the symmetric Kullback-Leibler divergence between $y_i$ and $z_t$, i.e.,

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

The parameters of KL-HMM (multinomial distributions) are trained using Viterbi expectation maximization algorithm with a cost function based on KL-divergence. In our studies, the KL-HMM is trained using the 2,880 training utterances of the RM database. The decoding is performed using standard Viterbi decoder. The reader may refer to [9] for more details on KL-HMM.

We compare the different features on both context-independent subword unit based system and word internal context-dependent subword unit based system. Each unit is modeled by a three state left-to-right HMM. The tuning parameters such as insertion penalty and language scaling factor were optimized on the development set.

## IV. SPEECH RECOGNITION STUDIES

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 II, for context-independent and context-dependent systems. In general, context-dependent systems show a reduction in WER over context-independent systems.

With MLP, *BBF* and *MF-PLP* perform comparably well, with WERs ranging from 5.1 to 5.6% for context-dependent,

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

| 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 II
WORD ERROR RATE (%) ON EVALUATION SET OF RM DATABASE USING CONTEXT-INDEPENDENT AND CONTEXT-DEPENDENT SUBWORD UNIT BASED SYSTEMS.

and 7.1 to 7.8% for context-independent. As reported in [12], 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*.

*BBF-TIMIT* and *BBF-RM* show similar performance. This suggests that *BBF* is not sensitive to the training data used for boosting, and can generalize well to unseen data.

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.

*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.

Overall, the performance of different features on RM corpus in terms of WER (Table II) shows similar trends as the frame accuracy results on RM corpus (Table I) and previous phoneme recognition results on TIMIT corpus [4].

## V. DISCUSSION AND CONCLUSIONS

This work investigated the use of Boosted Binary Features (*BBF*) for continuous speech recognition. Using MLP, *BBF* achieved comparable performance as standard cepstral features. Using SLP, binary features performed significantly better than cepstral features. It was found that the choice of data used for boosting the features was not critical and *BBF* could generalize well on unseen data.

The extraction of binary features could be interpreted as adding another layer to the MLP or SLP to learn phone-specific representations directly from the spectro-temporal plane using auxiliary data. This could have the potential to complement deep-learning frameworks geared towards similar objectives [14].

Possible directions for future work are outlined below:

1) *BBF* features were partly motivated by similar features proposed by the authors for speaker verification [15], which showed better noise-robustness compared to cepstral features. It would be interesting to verify this noise-robust characteristic for ASR also.
2) As *BBF* involves specific time-frequency points in the spectro-temporal matrix, it has potential to be directly coupled with suitable time-frequency masking frameworks (e.g. [16]) for noise removal or signal separation.
3) *BBF* are discrete-valued and has performed well with SLP. This indicates that they may be suitably incorporated into simpler modeling frameworks like Conditional Random Fields [17] with binary feature functions, instead of MLP followed by KL-HMM as in this work.
4) Since the extraction process of the parts-based binary features is distinct from standard cepstral features, they might have potential to capture complementary phoneme-specific information. Hence, fusion of these two features could improve performance.

## ACKNOWLEDGMENT

The authors would like to thank the Swiss National Science Foundation, project MULTI (MultiModal Interaction and MultiMedia Data Mining 200020-122062) and the FP7 European Research program, project MOBIO (IST-214324), for their financial support. The authors would like to thank Ramya Rasipuram for helping with MF-PLP experiments using MLP.

## REFERENCES

[1] S. Sharma and H. Hermansky, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. of ICASSP*, 1999.
[2] H. Hermansky and P. Fousek, "Multi-Resolution RASTA Filtering for Tandem based ASR," *Proc. of Interspeech*, pp. 361–364, 2005.
[3] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized Spectro-Temporal Cepstral Analysis of Speech," in *Proc. of ICASSP*, 2008.
[4] A. Roy, M. Magimai-Doss, and S. Marcel, "Phoneme Recognition using Boosted Binary Features," in *Proc. of ICASSP*, 2011.
[5] K. T. Schutte, "Parts-based Models and Local Features for Automatic Speech Recognition," Massachussets Institute of Technology, PhD Thesis, Jun. 2009.
[6] 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.
[7] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of CVPR*, 2001, pp. 511–518.
[8] J. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," *Annals of Statistics*, vol. 28, p. 2000, 1998.
[9] G. Aradilla, "Acoustic models for posterior features in speech recognition," Ph.D. dissertation, Ecole Polytechnique Fédérale de Lausanne, Lausanne , Switzerland, 2008.
[10] Y. Rodriguez, "Face Detection and Verification using Local Binary Patterns," Ecole Polytechnique Federale de Lausanne, PhD Thesis 3681, 2006.
[11] P. Price, W. Fisher, J. Bernstein, and D. Pallett, "The DARPA 1000-word resource management database for continuous speech recognition," in *Proc. of ICASSP*, 1988.
[12] J. Dines and M. Magimai.-Doss, "A study of phoneme and grapheme based context-dependent ASR systems," in *MLMI 2007, Lecture Notes in Computer Science*, no. 4892, 2008.
[13] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," *Proc. of DARPA Workshop on Speech Recognition*, pp. 93–99, Feb. 1986.
[14] A. Mohamed, G. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief network," *IEEE Trans. on Audio, Speech, and Language Processing (in press)*, 2011.
[15] A. Roy, M. Magimai-Doss, and S. Marcel, "Boosted binary features for noise-robust speaker verification," in *Proc. of ICASSP*, 2010, pp. 4442–4445.
[16] G. Lathoud, M. Magimai-Doss, B. Mesot, and H. Bourlard, "Unsupervised Spectral Subtraction for Noise-Robust ASR," in *Proc. of the ASRU Workshop*, 2005.
[17] J. Morris and E. Fossler-Lusier, "Conditional Random Fields for Integrating Local Discriminative Classifiers," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 16, no. 3, 2008.