USING BURST ONSET INFORMATION TO IMPROVE STOP/AFFRICATE PHONE RECOGNITION

Chi-Yueh Lin and Hsiao-Chuan Wang

Department of Electrical Engineering, National Tsing Hua University, Taiwan
bancolin@gmail.com, hcwang@ee.nthu.edu.tw

ABSTRACT

Reliably detecting salient phonetic-acoustic cues plays an important role in speech recognition based on speech landmarks. Once these speech landmarks are located, not only phone recognition can be performed but some other useful information can be derived as well. This paper focuses on the topic of detecting burst onset landmark, an important phonetic characteristic in stops and affricates. The proposed burst onset detector is based on random forest, a learning algorithm renowned for its high accuracy and efficiency in classification. By appending intermediate detection results to MFCCs, the expanded feature can bring benefit to the recognition of stop and affricate consonants in continuous speech.

Index Terms—random forest, phone recognition, stop consonant, affricate consonant, burst onset

1. INTRODUCTION

Burst onset is one of the most crucial speech landmarks in speech signal. Take stop consonants for example, several surveys approved that burst onset conveys many useful phonetic information [1, 2]. One can determine the place of articulation of a stop by inspecting the spectral shape of burst onset. Another major phonetic attribute, voice onset time (VOT), can also be derived if we can locate burst onset. However, burst onset could be the most transient landmark in speech signal. Unlike other sustained phonetic landmarks, a burst onset only lasts a few milliseconds. Conventional analysis configurations, e.g. a frame length of 25 ms and a frame shift of 10 ms, are not adequate to locate them within continuous speech. Typically a shorter frame length and a higher frame rate are needed when dealing with burst onsets.

From a spectrogram point of view, vertical stripes attributed to the burst onsets provide a useful cue to discriminate them from other speech landmarks. This sudden appearance in all-band energy during such a short period of time is the landmark we are looking for. Such acoustic behavior also inspires us to design a detector that can capture large energy variation in a burst onset region. In this paper we present the use of two-dimensional cepstrum (TDC) method [3] to capture the abrupt change around a burst onset. In our implementation, two-dimensional discrete cosine transform (2D-DCT) is exploited to accomplish the computation of TDC since 2D-DCT has the property of energy compaction. By combining the recent random forest (RF) technique [4], a RF-based detector is able to efficiently locate burst onsets within continuous speech.

The organization of the paper is summarized as follows. Section 2 introduces how a RF-based detector is constructed. Section 3 illustrates how the feature vector is derived. Section 4 gives the detail how the detection is performed. Then the final two sections give the experimental results and conclusion.

2. BURST ONSET DETECTOR

Breiman proposed random forest [4] by being inspired by several popular approaches: classification and regression tree (CART), bootstrapping and aggregating (bagging), and random subspace. The broad outline of a random forest is that it consists of many decision trees and the tree number can be up to hundreds or even thousands. The mechanism of a random forest is similar to a simple decision tree, but a random forest possesses an additional voting scheme at the final decision making stage. To train each decision tree in a random forest, randomness is not only injected in training sample selection, but also in node splitting phase. Here the latter randomization is introduced first, and the former one is left to the following paragraphs. To determine the optimal splitting hyperplane of an internal node in a D-dimensional space, we only consider d randomly chosen dimensions, where \(d << D\). This kind of node splitting is executed until all the leaf nodes achieve their highest purities, even leaf node only contains one training vector. The decision making phase of a random forest is much simpler, a testing feature vector is put through every decision tree, and its corresponding class label is determined by a plurality vote.

Before the construction of detector, we categorize 61 English phones in TIMIT dataset into six broad phonetic classes: stop, affricate, fricative, nasal, semivowel/glide, vowel, and non-speech. The desired burst onset landmarks are mainly extracted from the stop class. Unlike conventional modeling technique, where a whole stop segment is considered, we only concern about the burst onset portion. Thus we divide a stop segment into a burst onset portion and an optional remaining portion. The burst onset portion covers the closure-burst transition region where an abrupt energy change occurs, and it spans from 10 ms before the burst onset to 10 ms afterwards. If there is any portion that is not covered by the burst onset portion, it will be classified into the remaining portion.

However, there is one difficulty that we need to overcome before the random forest construction. Due to an uneven distribution of training vectors in broad phonetic classes, training a detector directly might result in a defective forest. The problem comes from the fact that our target class, stop, does not have as many training vectors as other classes. Using a conventional bootstrapping may have a chance not to select vectors from the target class, thus the resulting tree loses ability to identify stop class. To overcome the scenario of such uneven distribution, an
asymmetric bootstrapping is proposed to balance the training samples from each broad phonetic class. The asymmetric bootstrapping regards an uneven distributed training set as an input, and output an even distributed training set. The principle of asymmetric bootstrapping is that it selects training data from stop class first, and then from each other class. This procedure executes in an iterative fashion, says $T_{stop}$ iterations. The detail is summarized in Figure 1.

### 3. FEATURE REPRESENTATION

Speech landmark detection methods without using a segment-based model usually need to consider several frames simultaneously. Jayan et al. [5] grouped ten frames of broadband spectra and derived a rate-of-change function over this 10 ms interval. The strong, sharp spectral variation indicates a possible presence of a burst onset. Juneya et al. [6] encoded stop consonants in some knowledge-based acoustic parameters. Energy profiles such as energy onset and offset from a closure-burst transition are captured along a temporal patch spanned by ten consecutive frames. Niyogi et al. [7] derived three energy profiles featuring full-band, high-band, and the Wiener entropy of log-magnitude spectra and derived a rate-of-change function over this 10 ms interval. The strong, sharp spectral variation indicates a possible presence of a burst onset. Juneya et al. [6] encoded stop consonants in some knowledge-based acoustic parameters. Energy profiles such as energy onset and offset from a closure-burst transition are captured along a temporal patch spanned by ten consecutive frames. Niyogi et al. [7] derived three energy profiles featuring full-band, high-band, and the Wiener entropy of log-magnitude spectra and derived a rate-of-change function over this 10 ms interval. The strong, sharp spectral variation indicates a possible presence of a burst onset. Juneya et al. [6] encoded stop consonants in some knowledge-based acoustic parameters. Energy profiles such as energy onset and offset from a closure-burst transition are captured along a temporal patch spanned by ten consecutive frames. Niyogi et al. [7] derived three energy profiles featuring full-band, high-band, and the Wiener entropy of log-magnitude spectra and derived a rate-of-change function over this 10 ms interval. The strong, sharp spectral variation indicates a possible presence of a burst onset.

To encode a closure-burst transition, we choose the two-dimensional cepstrum (TDC) method [3] as our feature extraction procedure. In TDC, two Fourier transforms are performed successively on a spectro-temporal patch. The first Fourier transform applied on a log-magnitude spectrum is the same as the one in real cepstrum analysis. The derived cepstra are further analyzed by the second Fourier transform along the time axis to show their temporal variations. Apparently TDC formulates static and dynamic features under the same framework. Since log-magnitude spectra are real and even, cosine transform can be used to replace Fourier transform in the calculation [8]. Thus the most widely used transform in image processing, the two-dimensional cosine transform (2D-DCT), is adopted here.

Prior to TDC method, each utterance is processed by a 1-0.97 $z^{-1}$ pre-emphasis filter, and then is analyzed by a 24-order linear prediction (LP) analysis with a sliding Hamming window of length 10 ms (160 samples). The frame shift is chosen to be 2 ms (32 samples) to be capable of capturing burst onset in a high temporal resolution. Each frame is then transformed to LP spectrum by using a 512-point FFT. Once LP spectra are obtained, every ten contiguous frames are grouped into a single matrix so that the 2D-DCT can be applied on it to derive its corresponding TDC. Unlike the feature composition suggested in [8], where the first three columns of TDC matrix are concatenated as a feature vector, we prefer to extract coefficients from an isoceles triangular region, as one did in JPEG compression. Thus the dimension of a feature vector $x(t)$ is $55 (10^4(10+1)/2)$.

### 4. BURST ONSET DETECTION

During the detection phase a random forest detector $F$ votes for each input vector $x(t) \in \mathbb{R}^D$ as

$$v(t) = F(x(t), \Theta)$$

where $v(t) \in \mathbb{R}^C$ records votes for each phonetic class $c$. As mentioned in the previous sections, the dimension of feature vector $D$ is 55 and the number of phonetic classes $C$ is six. Eqn. (1) can be rewritten in terms of posterior probability by dividing the forest size,

$$p(c \mid x(t)) = \frac{v(t)}{|F|}$$

From $p(c \mid x(t))$ one can determine the most probable phonetic class at each time $t$,

$$\hat{c}(t) = \arg \max_c p(c \mid x(t)), \quad t = 1, \ldots, T$$

In order to verify whether or not the detected burst onsets (those $\hat{c}(t) = \text{stop}$) are prominent enough, a log-likelihood ratio score $\eta(t)$ is defined as follows,

$$\eta(t) = \log \frac{\max_{c \neq \text{stop}} p(c \mid x(t))}{\max_{c \neq \text{stop}} p(c \mid x(t))} \mathbb{I}(\hat{c}(t) = \text{stop}), \quad t = 1, \ldots, T$$

where $\mathbb{I}()$ is an indicator function,

$$\mathbb{I}(A) = \begin{cases} 1, & \text{if event } A \text{ is true} \\ 0, & \text{if event } A \text{ is false} \end{cases}$$

Eeq. (4) is positive when the most probable phonetic class $\hat{c}(t)$ is stop, and is zero otherwise. We further partition $\{\eta(t) \mid t \in \mathbb{I}\}$ into several homogeneous segments so that $\{\eta(t)\}$ in the same segment are either all zero or all positive. Let $S = \{s_1, s_2, \ldots, s_Z\}$ be a Z-segmentation of a time sequence $\{1,2,\ldots,T\}$, by which a sequence of length $T$ can be partitioned into $Z$ non-overlapped contiguous segments. For notation brevity, $\eta(s_c)$ denotes a sequence of $\{\eta(t) \mid t \in s_c\}$ of length $|s_c|$, and all $\eta(t)$ in a segment $\eta(s_c)$ are either all positive (burst onset presence) or all zero (burst onset absence). We also define an accumulated function $\Gamma(s_c)$ for each segment $s_c$,

$$\Gamma(s_c) = \sum_{t \in s_c} \eta(t)$$

Obviously, $\Gamma(s_c) > 0$ for burst onset presence, and $\Gamma(s_c) = 0$ for burst onset absence.

Similar to other verification tasks, each positive $\Gamma(s_c)$ will be compared to a positive threshold $\lambda$. The larger $\Gamma(s_c)$ is, the more confidence we have that there is a strong burst onset occurred in $s_c$. Here the threshold $\lambda$ is adjusted to a value $\lambda^{\text{EER}}$ so that the
detector presents its performance in terms of an equal error rate (ERR). To find out a detected burst onset, we check whether or not there is a segment \( s_z \) with \( \Gamma(s_z) > \lambda_{\text{EER}}^{\text{RF}} \) in a 10 ms neighborhood of a true burst onset, i.e., the manually labeled starting point of a stop or affricate segment.

5. EXPERIMENTAL RESULTS

5.1. Database

The performance of burst onset detection was evaluated on the TIMIT database. Owing to the non-parametric nature of random forest, training with a large number of training vectors will result in a huge random forest. To avoid an oversized random forest, the training data is limited to four speakers (two male and two female) from dialect region 1 (DR1). Within the selected data there are 71 voiced stop, 103 voiceless stop, and 1317 other phonetic segments. Training a random forest with the whole DR1 training data is feasible; however, the overall performance is only slightly better than that of the one with only four speakers. The testing data, on the other hand, uses all 1,680 utterances in the testing set, which in total with 6,991 stops, 631 affricates, and 56,522 other phonetic segments. Notice that in the testing set we do not exclude SA utterances.

With training data gathered from the selected four speakers, a 30-tree random forest is constructed. The number of iterations in the asymmetric bootstrapping, \( T_{\text{asy}} \), is set to 5. The randomly chosen dimensions, \( d_z \), in splitting the internal node, is set to 8, which is the nearest integer no less than the square root of \( D \).

5.2. General detection results

Each testing utterance is transformed to a sequence of feature vector \( \{ x(t) \} \) by the procedure mentioned in Section 3. Then the RF detector \( F \) regards these feature vectors as input and then outputs the presence of burst onsets according to whether or not \( \Gamma(s_z) \) surpasses the threshold \( \lambda_{\text{EER}}^{\text{RF}} \). In the performance assessment, burst onset presence due to the symbol /q/, a glottal stop or an irregular pitch period, is ignored. Generally speaking, the proposed RF-based burst onset detector achieves an EER of 7.33%. Since the true number of burst onsets is 7,623 (6,991 stops + 632 affricates), the missed detection count of true burst onsets is roughly 559 (7,623*0.0733), and the false alarm count is roughly 632 (7,623*0.0733). Comparing to other burst onset detection works \([5][7]\), the number reported here is comparable.

After examining the detail of detection results we found that a large proportion of false alarms are due to fricatives, e.g. \{/dh/, /hl/, /l/, /sl/, /sh/, /sz\}. Among those fricatives, dental fricative consonants /dh/ and /lh/ often induce a large value of \( \Gamma(s_z) \) so that half of them are erroneously being detected as burst onset presence. The reason is that their acoustic characteristics tend to be stop-like under some context. This finding also agrees with Zhao’s \([9]\) investigation on dental fricatives. Another major source of false alarms is from those silent segments /h/ that are blemished by microphone hissing sounds.

As for missed detections, most of them are due to weak burst onsets. Their closure-burst transition landmarks are not salient enough to be detected. It should be pointed out that some missed detections are due to erroneous reference boundaries. In a correctly manual-labeled stop segment the burst onset should lie between the two boundaries, however, this is not true in those erroneous manual-labeled cases.

5.3. Comparing to other learning machines

In this section two other popular learning machines will be used to detect burst onsets, and their detection performance will be compared to the proposed RF-based detector. The first one is support vector machine (SVM), and the second one is Gaussian mixture model (GMM). To be a fair comparison, all the training and testing data are the same for three learning methods. From the experimental results we find these three methods have their respective advantages and disadvantages on stop burst detection.

Two SVM-based detectors were built: the first one used radial basis function as kernel function (SVM-RBF), and the second one used linear kernel (SVM-LIN). To find optimal model parameters, e.g. slack variable \( C \) and radial width \( \gamma \), five-fold cross validations were performed in a grid search fashion. In SVM-RBF, the optimal parameters are \( C = 4 \) and \( \gamma = 0.25 \); and the optimal parameter in SVM-LIN is \( C = 1 \). All SVM training tasks were done with LibSVM toolkit. Figure 2 shows that SVM-RBF outperforms SVM-LIN in terms of EER (7.74% vs. 7.86%), but it is slightly worse than RF.

For GMM detectors, different mixture components were tested to find the best result. The numbers of mixture components used in the experiments were 4, 8, and 16. The GMM training tasks were done with Netlab toolbox. As shown in Figure 2, the detection performance keeps improving as the number of mixture components is increasing. The best GMM detector is the one with 16 mixture components, achieving 8.79% in EER. Using more mixture components, on the contrary, suffers from the over-fitting problem because the training vectors are not enough to robustly estimate GMM parameters.

In conclusion, when the detection performance in terms of EER is concerned, RF slightly outperforms SVM-RBF and SVM-LIN by 6% relatively. GMM-based detectors, however, perform about 10%~20% worse than the previous two learning methods. Regarding the detection performance, both RF and SVM have advantages over GMM. When computation time is concerned, RF is also stands out among the three due to its tree-structured nature. GMM follows RF as the second efficient learning algorithm. SVM needs more computation time during the training and testing.

Figure 2. Performance comparison of burst onset detection using different learning algorithms.
5.4. Inclusion of burst onset information in phone recognition

In order to examine whether or not the burst onset information (BOI) brings benefit to phone recognition, we regarded the burst onset detection results as auxiliary information and appended them to the conventional MFCC vectors. It is expected that those phones possessing burst onset landmarks would be benefitted in phone recognition. The phones to be examined are six stops (/b/, /d/, /g/, /p/, /t/, and /k/) and two affricates (/ch/ and /jh/). Since a modern speech recognizer commonly uses 10 ms as frame shift, which is incompatible to the 2 ms frame shift in our burst onset detector, a simple averaging strategy is applied on the burst onset detection results before they are appended to MFCC feature vectors.

Owing to the fact that the burst onset detection information is appended in the feature level, some pre-processing must be done upon the training and testing data. At first the RF detector detected burst onsets in all training utterances. The intermediate information, \( p(c = \text{stop } x(t)) \) and \( p(c = \text{fricative } x(t)) \) in Eqn. (2), is regarded as two extra dimensions and appended to a MFCC vector. The reason of including fricative information is that both affricates and aspirated stops have a portion of fricative segment. The same manipulation is done in the testing utterances as well.

Our baseline phone recognizer was trained on TIMIT database by using HTK. Unlike the previous burst onset detection, here SA sentences are excluded from training and testing. The recognition setting followed Lee's work [10]: 48 monophone HMMs were built and during the recognition phase they were further mapped to 39 phones. Each monophone HMM consists of three states and follows a left-to-right topology. Each feature vector is composed of 44 dimensions: 13 static MFCCs and 1 static log-energy, their corresponding delta, delta-delta coefficients, and the two additional dimensions produced by the detector. To simplify the training process, the number of mixture components in each state was fixed to eight for every monophone HMM, and no other model refinement strategies were applied. During the recognition phase no language models were applied, so the recognition results were purely based on acoustic models. In this fashion our phone recognizer looked quite primitive, but it was sufficient to reflect the impacts brought by any modification of acoustic models.

Figure 3 shows the recognition rates with and without including BOI. BO denotes the average rate of eight phones with burst onset, Non-BO denotes the average rate of other 30 phones, and Total is the average rate of all 38 phones. Note that the rate of silence model was excluded in this presentation. When the correction rate is concerned, only /g/ has an adverse effect, as its rate degrades slightly. The correction rate of Non-BO is also decreased because the included BOI doesn’t favor these phones. As far as accuracy rate is concerned, BO still has an improved performance, except that velar stops (/g, k/) and /b/ have a performance drop due to an increase of insertion errors. The effect of including BOI to other phones is limited.

6. CONCLUSION

This paper investigates the topics on burst onset detection and its possibility to improve the phone recognition. A burst onset detector based on random forest can effectively locate burst onsets within continuous speech, and its detection performance also outperforms those detectors based on SVM and GMM. Regarding the detection results as auxiliary information and appending them to the conventional MFCC vector, the resulting feature vectors can bring benefit to those phones possessing burst onset landmarks.

Acknowledgement

This research was sponsored by the National Science Council, Taiwan, under contract number NSC-97-2221-E-007-075-MY3

REFERENCES