Application of Local Binary Patterns for SVM based Stop Consonant Detection

Kaizhi Qian, Yang Zhang, Mark Hasegawa-Johnson

Univeristy of Illinois at Urbana-Champaign, USA

kqian3@illinois.edu, yzhan143@illinois.edu, jhasegaw@illinois.edu

Abstract

Detection of acoustic phonetic landmarks is useful for a variety of speech processing applications such as automatic speech recognition.The majority of existing methods use Mel-frequency Cepstral Coefficients (MFCCs) describing the short time power spectral envelope of the speech signal. This paper hypothesizes that a different feature extraction method can be used to complement MFCCs by capturing more complex transient acoustic cues. The proposed feature extraction method quantizes spectrogram textures using local binary patterns (LBP). This paper particularly exploits landmark based stop consonant detection. Both methods outperform the previous work on stop consonant detection and the latter is particularly appealing for real time detection in which computation efficiency matters.

Index Terms: acoustic phonetic landmark detection, stop consonants, local binary pattern, time-frequency features

1. Introduction

An acoustic phonetic landmark is a perceptually salient instantaneous acoustic event caused by an extremum or discontinuity in the pattern of airflow through the mouth [1]. It is claimed a consonant closure followed by a consonant release is sufficient to determine the consonant being uttered [2]. According to [3], automatic speech recognition can be decomposed into a collection of landmark detection tasks. In the speech synthesis literature, landmark detection is called "phone segmentation" and is an essential pre-requisite for good quality concatenative synthesis [4]. Scientific studies of the acoustic correlates of prosody [5] and rhythm [6], similarly, depend on consonant/vowel segmentation as a pre-requisite. The study of stop consonant detection is particularly interesting because it is one of the phone labeling tasks performed by human transcribers with highest accuracy, e.g., with 99% precision and 99% recall at 6dB SNR [7]. Once detected, a stop consonant can be classified by a neural net with 98.5% accuracy [8], but the best published figures for automatic stop consonant detection from untranscribed speech still have unreasonably high equal error rates (e.g., 16.5% [6]).

The selection of appropriate features is crucial to the performance of detection. A good feature produces representations that distinguish events from non-events using as few dimensions as possible. For phonetic landmark detection, manually selected features followed by support vector machine have shown reasonable results. In particular, a three dimensional feature vector consisting of log of total energy, log of energy above 3 kHz and spectral flatness is computed for each frame for stop consonant detection [6]. However, the manually selected features are difficult to generalize to other phonetic landmarks, because the choice of acoustic features varies with respect to the type of phonetic landmarks. Recently, landmark detection is primarily based on mel-frequency cepstral coefficients (MFCCs) [9] [10]. MFCCs capture the shape of the vocal tract by computing the short-time spectral envelope. MFCCs are computed over each frame of the speech signal, during which the signal is assumed to be stationary. However, MFCC features may not be sufficiently discriminative for phonetic landmarks because landmarks are transient non-stationary acoustic events.

In order to overcome the above disadvantages, this paper describes a new type of feature characterizing the acoustic structure from the texture feature of speech spectrograms. A speech signal can be transformed into a two-dimensional time-frequency representation as a spectrogram, upon which image processing techniques can be applied. Spectrogram based audio signal processing gave favorable results for acoustic scenario detection and sound classification [11] [12] [13]. The proposed feature is motivated from an image object detection technique, the local binary pattern (LBP), to capture the transient change of phonetic landmark as an object on the spectrogram of the speech signal. LBP along with its variants are highly discriminative and compact for image texture classification. However, the classic LBP is highly affected by the fluctuation of pixel values seen in the spectrogram. Thus, the proposed method thresholds the pixels using the mean of the local patch and weights the corresponding frequency counts using the standard deviation of the local patch in LBP computation in order to mitigate the fluctuation of pixels values. Meanwhile, the dimension is reduced by considering only the uniform LBP to further reduce the effect of noisy pixels. Finally, L2-Hellinger normalization is applied to the LBP feature to make it more discriminative for SVM [14] classification using a linear kernel.

In this paper, we particularly test the proposed method for stop consonant detection, because stop consonants display distinctive transient acoustic cues that are perceptually salient during speech recognition. LBP will be able to handle most other landmarks if it can successfully capture the rapid acoustic structural change of a stop closure followed by stop release. The proposed feature extraction method is tested on the TIMIT corpus that contains 14000 stop consonant audio clips. The proposed LBP shows a small non-significant accuracy gain relative to the MFCC baseline, at significantly lower computational cost. More importantly, it provides useful insights in how stop consonants are discriminated from other phones acoustically.

2. Improved local binary pattern

The local binary pattern (LBP) has been a discriminative and compact image texture feature representation for visual object detection and classification [15]. By regarding the time-frequency spectrogram as an image, LBP is able to capture the transient temporal and frequency variations of the landmark as the gray level change of pixels across the time dimension and frequency dimension respectively.

The original LBP is defined in a 3×3 local patch centered at each pixel of the region of image being analyzed, where the neighboring 8 pixel values are thresholded using the center pixel value. The location of the pixel being compared is set to 1 or 0 depending on if it is higher or lower than the center pixel value. The 3×3 binary pattern is then unwrapped clock-wisely around the center pixel into a binary bit string, which can be mapped to a decimal number whose respective frequency counts are summarized in a histogram. The major problem with the original LBP is that it is easily affected by pixel value fluctuation. In particular, the original LBP only takes into account the relative magnitude of pixel values without considering the actual difference between them, which makes the pixels having close pixel values vulnerable to pixel fluctuations. If one pixel is slightly lower than the threshold, a small amount of fluctuation can flip the corresponding bit from 0 to 1, which breaks up the local pattern. As a result, patches distorted by noise with a small variance can be mistakenly mapped to the incorrect local patterns.

Statistics based uniform LBP [11] is employed in this paper for enhancing the noise robustness and reducing the dimensionality of the original LBP. First, the mean pixel value of the local patch is used instead of the center pixel value to threshold the neighboring pixel values including the center pixel value, because the mean pixel value is relatively insensitive to pixel value fluctuations. Let g_i and μ_c be the i th pixel value and the mean pixels value of the local patch respectively. The binary pattern is calculated and converted to a decimal index as

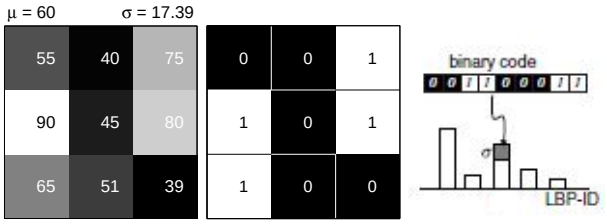
$$LBP_{T,F} = \sum_{i=0}^{T \cdot F - 1} f(g_i - \mu_c) 2^i, \quad f(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0 \end{cases} \quad (1)$$

where g_i is unwrapped clock-wisely as shown in Figure 1c, and T and F are the dimension of the local patch along the time axis and frequency axis respectively. Equation 1 generates $2^{T \cdot F}$ binary patterns corresponding to decimal indexes in the range from 0 to $2^{T \cdot F} - 1$ whose respective frequencies are counted into the histogram. Second, the histogram accumulates the standard deviation of the local patch instead of the frequency count for its corresponding binary local pattern. In particular, the local patch scans over each pixel in region A of the image being analyzed and accumulates the standard deviation of the patch, $\sigma_c = \sqrt{\sum_{i=0}^{T \cdot F - 1} (g_i - \mu_c)^2}$ into the corresponding decimal indexed location of the histogram as expressed in Equation 2

$$x_i = \sum_{c \in A} I_{\{LBP^{dec}_{T,F} = i-1\}} \sigma_c, \quad i = 1, \cdots, 2^{T \cdot F} \quad (2)$$

where $LBP^{dec}_{T,F}$ is the decimal value of the binary pattern, i starting from 1 is the decimal index of the corresponding uniform binary pattern in the histogram, c is the center pixel of the local patch, and $I_{\{\cdot\}}$ is the indicator function that equals to 1 only if the equation in the brackets holds and to 0 otherwise. An example schematic is shown in Figure 1. The local binary patterns corresponding to patches with small standard deviations are vulnerable to pixel value fluctuations, thus their frequency counts in the histogram are scaled down by their small standard deviations. On the other hand, the binary patterns corresponding to patches with large standard deviation are considered more noise-robust and significant, thus their frequency counts in the histogram are emphasized by scaling up by their large standard deviations.

It is worth noticing that the number of different binary patterns grows exponentially with respect to the number of pixels



(a) local patch (b) binary pattern (c) weighted counts

Figure 1: *Each local patch (a) is encoded into a local binary pattern (b) and accumulates $\sigma = 31$ instead of 1 into the histogram (c) at the corresponding decimal index.*

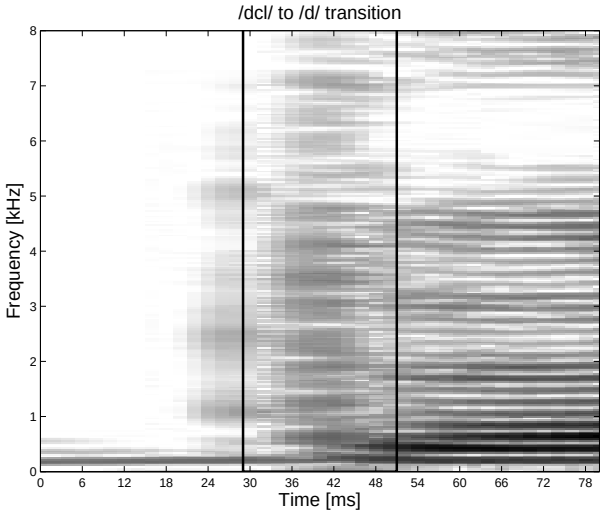


Figure 3: *Spetrogram of transition from /dcl/ to /d/ (20ms frames w/2ms shift). The boxed region is the extracted positive example.*

in the local patch. There are already 256 different binary patterns for the 8 pixel patch. Large number of different binary patterns results in higher dimensional and more sparse histograms, which is likely to cause over-fitting and be more vulnerable to noise. Ojala et al. [15] claim that the majority of the local binary patterns only have less than two 1-0 or 0-1 transitions when their respective binary bit string is circularly connected to itself. Such patterns are referred to as the uniform binary patterns that are most informative about the texture structure, while the non-uniform patterns are considered insignificant and not noise-robust for texture classification. Therefore, only uniform local binary patterns are represented by distinctive bins in the histogram, while the non-uniform binary patterns are grouped into one single bin in the histogram. Limitation to uniform binary patterns not only reduces the histogram dimension, but also improves the noise robustness of the pattern distribution. For an 8-bit patch, the histogram only has 59 bins, where the 1st bin to the 58th bin count the weighted frequency of the 58 uniform patterns, while the last bin counts all the non-uniform patterns. A lookup table from the original binary patterns to the uniform patterns is computed for mapping the bins of histogram.

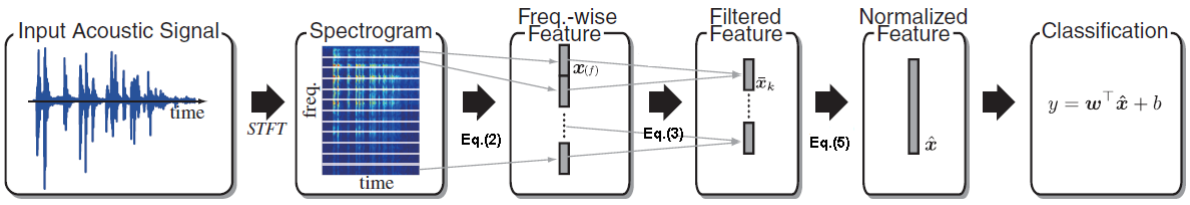


Figure 2: *Basic flow chart of LBP feature extraction*

3. Classifier architecture

Stop burst is produced by a stop closure followed by a stop release, and is the most distinctive characteristic to indicate the presence of a stop consonant [6]. Thus, any instant of a stop closure followed by a stop release with the boundary between them at the center is a positive example for the SVM. Everything else including vowels, fricatives, nasals, silences, etc. are considered as negative examples for the SVM. The following sequence shows a typical positive example[1],

$$\underset{t_{begin}}{ih-|-dcl-|-d-|-ah-|}\cdots \qquad \underset{t_{center}}{} \qquad \underset{t_{end}}{}$$

where the center time of the frame is at the boundary between stop closure and stop release. The begin time and the end time are equidistant from the center time to include a certain amount of context information. The spectrogram of the above utterance is shown in Figure 3. Each positive or negative example is cut from the speech signal according its phone level transcription.

3.1. Mel-frequency cepstral coefficient

As our baseline feature, 13 dimensional MFCC and its deltas and delta-deltas (39 dimensions in total) are extracted for each frame, with 20ms frame length and 2ms frame shift. We set the frame shift to 2ms because it is the time resolution of human auditory systems for broad band noise signals [18]. The MFCC features of 11 consecutive frames centered at the boundary between stop closure and stop release are concatenated to form a 39×11 matrix $[\bar{x}_{c-5}, \cdots \bar{x}_c \cdots, \bar{x}_{c+5}]$, where $\bar{x}_{c-5}$ and $\bar{x}_{c+5}$ are centered at $(t_{center} - 10)$ ms and $(t_{center} + 10)$ ms respectively. The 11 feature vectors are transposed and horizontally concatenated to produce a training or testing token given as a 1×429 row vector $\bar{x}_k = \left[\bar{x}_{c-5}^T, \cdots \bar{x}_c^T \cdots, \bar{x}_{c+5}^T\right]$. Finally, CMVN [19] is applied by subtracting the mean and dividing the standard deviation of each dimension.

We extracted 14000 positive and negative training tokens respectively from TIMIT. All the tokens form a 28000×429 MFCC feature matrix for the SVM.

3.2. Local binary pattern extraction

The real-valued time-frequency spectrogram is calculated using short time Fourier transform (STFT) of length 512 for each frame, with 20ms frame length and 2ms frame shift. Statistics based uniform LBP summarizes each frequency bin along the time axis into a sparse histogram $\bar{x}$. Related works [11] [12] have shown that local patch having around 8 pixels gives the best performance. We primarily exploit the 2×4 and 4×2 time-frequency patches, which produce 59-dimensional frequency-wise histograms using uniform binary patterns. Explicitly, the proposed feature extracts a histogram at each frequency bin, thus the area defined by $\mathbb{A}_f = \{a = (t', f') | \forall t', f' = f\}$ to produce a feature vector $\bar{x}(f)$ of length 59 at each frequency bin. It is worth mentioning that the local binary pattern distributions of two different temporal alignments of the same signal are completely different. Therefore, the global temporal variation is marginalized out so that each frequency-wise histogram is invariant to time shifts across utterances and only describes the local transient temporal variations of the stop consonant.

To better match the feature characteristics to the characteristics of human hearing and further reduce the frequency dimensionality, 30 triangular filters $H_k(f)$ spaced uniformly in an equivalent-rectangular-bandwidth (ERB) frequency scale [20], where k is the ERB frequency index and f is FFT bin number, are applied to the 257×59 feature matrix. Given the frequency response of each filter bank $H_k(f)$, the frequency-wise resampled feature matrix is calculated as

$$\bar{x}_k = \sum_f H_k(f)\bar{x}(f), \bar{x} = \left[\bar{x}_1^T, \cdots, \bar{x}_K^T\right]^T. \qquad (3)$$

Finally, the rows of the filtered feature matrix are horizontally concatenated into a feature vector of length 1770, after which the L2-Hellinger normalization [21] that is effective for measuring similarities between histograms is applied. Stacking all the tokens vertically gives a 28000×1770 token matrix. The basic flowchart is shown in Figure 2.

4. Experimental results

4.1. Overall Performance

We evaluate the performance of the proposed method on stop consonant detection using the TIMIT corpus. The TIMIT corpus contains 6300 sentences spoken by 630 different American English speakers. The 28000 training tokens are extracted from the full training set of TIMIT, and the 8000 testing tokens are extracted from the full testing set of TIMIT. The ROC curve for our baseline MFCC feature and our proposed LBP feature is shown in Figure 4. It can be seen that the LBP slightly outperforms the MFCC, especially in terms of the false negative rate. The computation time of training and testing using LBP is more than 10 times less than that of MFCC. A more comprehensive comparison table shows our proposed method has the state-of-art performance.

We test the performance for different patch sizes and the best performance is obtained on 2×4 (time $\times$ frequency). The performance is better for local patches with higher frequency dimension than time dimension. The most discriminative LBPs all show energy onset in the time dimension; a 2×4 patch can

[1]This paper uses TIMIT's ARPABET phoneme code [16] rather than IPA, because IPA does not distinguish the closure and release subsegments of a stop consonant.

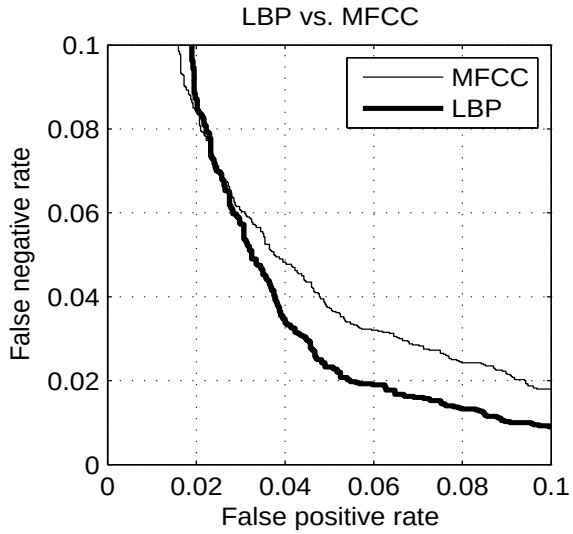


Figure 4: *The ROC curve for MFCC and LBP*

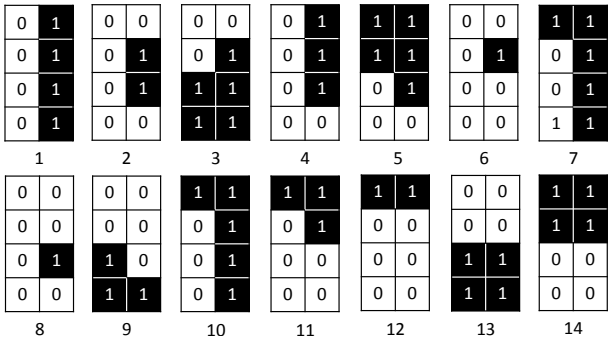
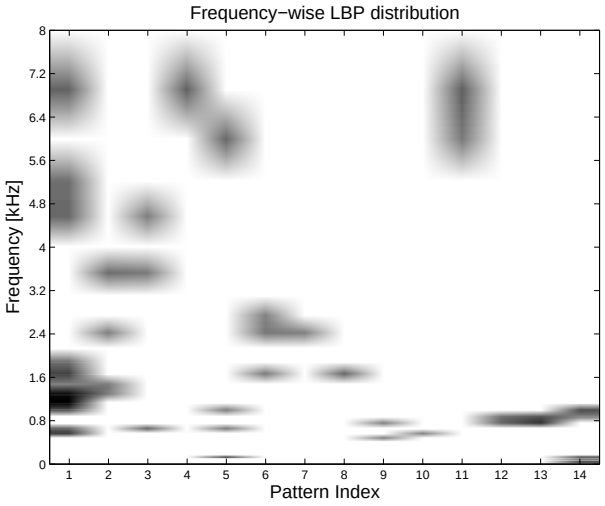


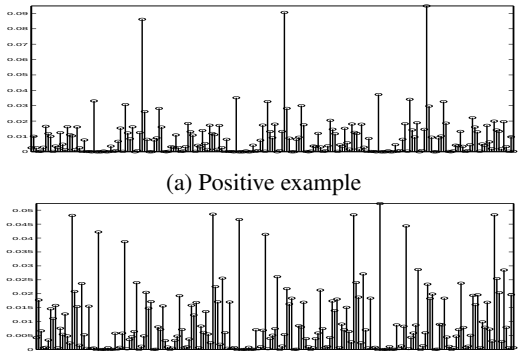
Figure 5: *The ROC curve for MFCC and LBP*

Table 1: *Equal Error Rate (EER) and area under ROC curve (AUC) of stop consonant detectors in this paper, and in other published work if applicable.*

Method	EER.%	AUC.%
Our MFCC	4.42	0.988
Our LBP	**3.83**	0.991
Borys' MFCC [9]	6.24	NA
Niyogi et al. [6]	16.50	NA

Figure 6: *The frequecny-wise distribution of the 15 most discriminative patterns. Darker gray level indicates the corresponding LBP is activated by the SVM with higher weight.*





(a) Positive example

(b) Negative example

Figure 7: *Portions of the feature vectors of positive and negative examples display different periodic patterns.*

represent frequency-dependent onset patterns, which is apparently more useful than the extra temporal information in a 4×2 patch.

4.2. The Sparse Pattern of the LBP Feature

Although the dimension of the proposed LBP feature is 1770. It turns out that it displays a very consistent sparse pattern, which greatly facilitates dimension reduction.

Figure 3 shows the average value of a portion LBP feature across all positive examples (upper panel) and negative examples (lower panel). As can be seen, both averaged features displays similar sparse patterns, i.e. the dimensions of very small values are similar. Therefore, an intuitive dimension reduction approach is to keep only the dimensions with value greater than a threshold. After removing the small spikes, the feature dimension drops from 1770 to 1051, which slightly improves the performance and greatly reduces computation time.

4.3. Discriminative Patterns

This subsection analyzes the most discriminative local time-frequency patterns, which gives us useful insights into the frequency and temporal characteristics of stop consonants. In SVM, the magnitude of the weights implies the importance of the corresponding feature dimension on the classification task. Recall that in the proposed LBP-based classifier, each feature dimension corresponds to a local pattern in a specific frequency bin. Therefore, we can evaluate the level of discrimination of each pattern by summing all the weights that correspond to this pattern across frequencies.

Figure 5 plots the 14 most important patterns ranked by their importance, and figure 6 plots their distribution of importance across frequencies. There are several useful findings.

First, we can observe a strong asymmetry along time axis. As time proceeds, we can see a non decreasing energy trend, namely the number of 1's in the right column is almost always no smaller than that in the left column, except for pattern 38. This reflects that one of the most important temporal characteristics of stop consonant is energy burst, i.e. the energy increases abruptly as time proceeds, which is a key distinctive feature of stop consonant and is used as the baseline for stop consonant detection by Niyogi et al. [6].

More specifically, patterns 1, 2, 4, 6, 7 and 8 best capture the energy burst pattern, and most of them rank very high. In particular, pattern 26, which ranks 1st in importance, is the most typical energy burst pattern. According to its frequency distribution of importance as shown in figure 6, there are 3 frequency bands where the energy burst is most significant: 0.8-2 kHz, 4-5.6 kHz, and 6-7.5 kHz, which agrees well with the findings in [22], which identifies the most discriminative frequency bands of stop consonants for human perception.

On the other hand, the patterns are quite symmetric along frequency axis. For example, pattern pair 3 and 5, 6 and 8, and 13 and 14 are completely symmetric along the frequency axis, and the rankings within a pair are quite close. This suggests that the rise and fall in energy along frequency axis is quite balanced. In particular, as frequency increases, patterns 5, 11 12 and 14 depict energy rises, and patterns 3, 9 and 13 depict energy fall. These patterns together describe the fine spectral structure. In terms of distribution in frequency, these patterns concentrate in low frequencies. One possible explanation is that in low frequency, the spectral structure of different phones, particularly voiced and unvoiced phones, differ more than in high frequency, where voiced energy is heavily corrupted with noise and aspiration. Hence, the spectral structure in low frequency is more discriminative.

5. Conclusions

In this paper, we exploit the discriminative power of the statistics based uniform local binary pattern features (LBP) for stop consonant detection. The LBP extracts acoustic phonetic structural change as texture feature by viewing the time-frequency spectrogram as an image. Our LBP-based method successfully extracts the discriminative features and shows superior performance for landmark-based stop consonant detection with high computation efficiency.

6. References

[1] K. N. Stevens, "Models of speech perception based on acoustic phonetic landmarks," in *Proc. Interspeech*, 2000.

[2] S. A. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," *J. Acoust. Soc. Am.*, vol. 124, no. 2, pp. 1220–1233, 2008.

[3] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a model for lexical access based on features," in *Internat. Conf. Spoken Lang. Process.*, vol. 1, Banff, Alberta, 1992, pp. 499–502.

[4] H. Kawai and T. Toda, "An evaluation of automatic phone segmentation for concatenative speech synthesis," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, vol. 1, May 2004, pp. I–677–80 vol.1.

[5] C. W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *The Journal of the Acoustical Society of America*, vol. 91, no. 3, pp. 1707–1717, 1992.

[6] P. Niyogi, C. Burges, and P. Ramesh, "Distinctive feature detection using support vector machines," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1, 1999, pp. 425–428.

[7] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[8] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 3, pp. 328–339, 1989.

[9] S. E. Borys, "An svm front-end landmark speech recognition system," Ph.D. dissertation, Citeseer, 2008.

[10] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan *et al.*, "Landmark-based speech recognition: Report of the 2004 johns hopkins summer workshop," in *Proceedings of the... IEEE International Conference on Acoustics, Speech, and Signal Processing/sponsored by the Institute of Electrical and Electronics Engineers Signal Processing Society. ICASSP*, vol. 1, no. 1415088. NIH Public Access, 2005, p. 1213.

[11] T. Kobayashi and J. Ye, "Acoustic feature extraction by statistics based local binary pattern for environmental sound classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3052–3056.

[12] D. Battaglino, L. Lepauloux, L. Pilati, and N. Evans, "Acoustic context recognition using local binary pattern codebooks," in *WASPAA 2015, IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 18-21 October 2015, New Paltz, NY, USA*, New Paltz, ÉTATS-UNIS, 10 2015. [Online]. Available: http://www.eurecom.fr/publication/4721

[13] C.-Y. Wang, Y.-H. Chin, T.-C. Tai, D. Gunawan, and J.-C. Wang, "Automatic recognition of audio event using dynamic local binary patterns," in *Consumer Electronics - Taiwan (ICCE-TW), 2015 IEEE International Conference on*, June 2015, pp. 246–247.

[14] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[15] T. Ojala, M. Pietikäinen, and T. Mäenpää, "A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification." in *ICAPR*, vol. 1. Springer, 2001, pp. 397–406.

[16] V. Zue, S. Seneff, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[17] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[18] R. Plomp, "Rate of decay of auditory sensation," *The Journal of the Acoustical Society of America*, vol. 36, no. 2, pp. 277–282, 1964.

[19] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1, pp. 133–147, 1998.

[20] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[21] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[22] F. Li and J. B. Allen, "Manipulation of consonants in natural speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 3, pp. 496–504, 2011.