A ROBUST SPEECH RATE ESTIMATION BASED ON THE ACTIVATION PROFILE FROM THE SELECTED ACOUSTIC UNIT DICTIONARY

Supriya Nagesh[1] *Chiranjeevi Yarra*[2] *Om D. Deshmukh*[3] *Prasanta Kumar Ghosh*[2]

[1]National Institute of Technology Karnataka (NITK), Surathkal-575025, India
[2]Electrical Engineering, Indian Institute of Science (IISc), Bangalore-560012, India
[3]Xerox Research Center India, Bangalore, India

ABSTRACT

A typical solution for the speech rate estimation consists of two stages, which involves first computing a short-time feature contour such that most of peaks of the contour correspond to the syllable nuclei followed by the detection of the peaks of the contour corresponding to the syllable nuclei. Temporal correlation selected sub-band correlation (TCSSBC) is often used as a feature contour for the speech rate estimation in which correlation within and across a few selected sub-band energies are computed. In this work, instead of a fixed set of sub-bands, we learn them in a data-driven manner using a dictionary learning approach. Similarly, instead of the energy contours, we use the activation profile from the learned dictionary elements. We found that the peaks detected from the data-driven approach significantly improve the speech rate estimation when combined with the traditional TCSSBC approach using a proposed peak-merging strategy. Experiments are performed separately using Switchboard, TIMIT and CTIMIT corpora. Except Switchboard, the correlation coefficient for the speech rate estimation using the proposed approach is found to be higher than those by the TCSSBC technique – 3.1% and 5.2% (relative) improvements for TIMIT and CTIMIT respectively.

1. INTRODUCTION

Automatic speech recognition (ASR) accuracy has been often shown to improve by using estimated speech rate [1] [2], which results in better human computer interface in the applications such as computer assisted language learning (CALL) and fluency analysis [3] [4]. Typically, speech rate is estimated by counting the number of speech units per second. Most of the existing works in the literature use syllable as the speech units [1] [5] [6]. The speech rate estimation usually involves identification of the syllable nuclei locations followed by syllable rate computation [7]. Generally the approaches for the speech rate estimation are based on either acoustic feature [1] [5] [6] [7] or ASR based recognition systems [3] [4] [8] [9]. The ASR based methods involve decoding the speech signal into phonetic/syllabic transcription, following by the speech rate estimation. However, ASR systems are not very reliable in the presence of noise as well as for spontaneous speech especially where reference transcription is not available [5]. Due to error in recognition, speech rate based ASR model would accumulate the errors in recognition. In such scenarios, acoustic feature based speech rate estimation is preferred [1] [2].

A typical acoustic feature based approach consists of two steps: 1) computing a short-time feature contour such that most of the peaks correspond to the syllable nuclei locations, 2) detecting the peaks in the contour belonging to the syllable nuclei. For example, Pfau et al. estimated the vowel locations based on prominent peak locations in the smoothed loudness contour [10]. A Hilbert-envelope-based contour was used by Zhang et al. to estimate the syllable nuclei [11]. Landsiedel et al. proposed a contour based on long-short-term-memory neural-networks [12]. Similarly, Jong et al. used intensity-based envelope with simple peak counting based on voicing decisions to estimate speaking rate [13]. Wang et al. introduced a method by proposing a feature contour "temporal correlation and selected sub-band correlation (TCSSBC)", which involves computing a spectral and temporal correlation [14]. As described by Dekens et al. [15], the TCSSBC achieves the highest speech rate estimation accuracy among other features.

The TCSSBC exploits formant-like structures by using 19 fixed sub-band energy profiles. Temporal correlation and spectral correlation (referred to as spectro-temporal correlation (STC)) are computed on the selected sub-band energies to maximize the peaky nature of the feature contour around the syllable nuclei. We hypothesize that instead of representing the spectral structures with a fixed set of sub-band energies, learning those structural information in a data-driven manner would be more effective. We observe that the TCSSBC contour is often not peaky at the syllable nuclei because the sub-band energy based representation is generic [16] in the sense that it does not explicitly utilize the spectral structure of the acoustics near the syllable nuclei. In contrast to a generic representation using a fixed set of sub-bands, we aim to obtain representation based on the spectral structure of selected acoustic units within syllable in a data-driven manner. For learning the structural information of speech spectra we use non-negative matrix factorization (NMF) in which a set of bases (called dictionary) is learnt. The NMF has the ability to learn the dictionary elements that capture intrinsic parts/structures in the high-dimensional data [17]. This property has motivated us to investigate usefulness of the NMF in the speech rate estimation. A typical approach of NMF involves two steps – 1) learning a dictionary matrix from the training data vectors that represents data structures, 2) representing new data vector sequence as a linear combination of the dictionary elements using a set of activation weights [18].

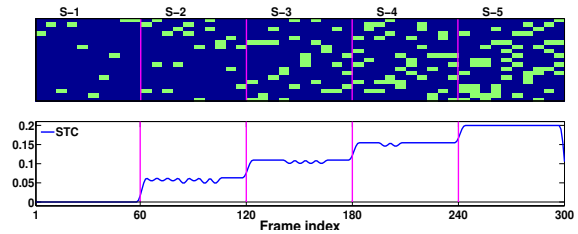

Fig. 1. Illustrative example for analyzing STC from synthetic matrices with different sparsity and consistency values

We, in this work, compute a feature contour based on STC using the activation matrix obtained from NMF. For a good quality speech rate estimation, the feature contour should have a peak in every syllable. This in turn requires that near the expected peak location of the feature contour, the activation matrix is less sparse in each column & highly consistent in each row and near the valleys of the feature contour, the activation matrix is sparse and less consistent. In Figure 1, we illustrate this observation using five synthetic matrices each spanning 60 columns and the feature contours obtained from a large matrix concatenating those five by applying STC (equation 1 and 2 proposed by wang et al. [14]). I^{th} synthesized matrix is denoted by S-i (shown above the top row in Figure 1), for which only i elements in every column is non-zero. With increasing i, i.e., decreasing sparsity, we synthesize the matrices such that they are more consistent in each row indicated by green horizontal stripes. These consistent patterns are clearly seen in S-5 matrix indicated by green horizontal stripes. The different values of STC for different matrices suggest that the dictionary should be learnt in such a way that the activation matrix for a test utterance is less sparse and more consistent near the expected peak location within a syllable and sparse and less consistent in the remaining portion of the syllable. This could result in peaky feature contour in the targeted portion of the syllable. A generic dictionary learnt from speech spectra of different sound categories is found to be inappropriate for this purpose.

We, in this work, consider learning dictionaries from frames of four different types of acoustic categories. We perform STC on the activation matrix for computing feature contour called NMF-TCSSBC. Peaks in the NMF-TCSSBC are detected by following the steps of peak detection strategy proposed by Wang et al. [5]. We propose a peak merging technique to combine the detected peaks from the NMF-TCSSBC with the detected peaks from the TCSSBC. The effectiveness of the proposed dictionary based (DB) approach is demonstrated using three large corpora, namely, Switchboard (SWBD) [19], TIMIT [20] and CTIMIT [21]. Experiments for the speech rate estimation are performed on each corpus separately. The proposed DB speech rate estimation achieves better performance in comparison to the TCSSBC.

2. DATABASE

We use ICSI SWBD [19], TIMIT [20] and CTIMIT [21] corpora for all experiments in this work. SWBD is a spontaneous speech corpus consisting of sentences spoken by 370 speakers with a wide range of speech rate, ranging from 1.26 to 9.2 syllables per second. The audio in the SWBD corpus was collected through the telephone channel. A subset of 7300 audio segments, each of duration greater than 200ms, is used for our experiments. TIMIT is a read speech database, which has phonetically balanced 6300 sentences spoken by 630 speakers with a speech rate ranging from 1.44 to 8 syllables per second. All sentences from the TIMIT are used for our experiments. CTIMIT corpus is similar to TIMIT except that the audio was collected through the cell phone channel under various noisy conditions. All 3370 sentences from the CTIMIT, spoken by 630 speakers, are used for our experiments. The speech rate in the CTIMIT sentences ranges from 1.87 to 8 syllables per second. Using the available phonetic transcriptions, silent segments in the initial and final parts of each sentence of all corpora are removed.

3. PROPOSED DICTIONARY BASED APPROACH

The steps involved in the proposed DB speech rate estimation are described with the help of a block diagram in Figure 2. The block diagram has two major stages – a) feature contour (NMF-TCSSBC) computation; b) merging based peak estimation. The feature computation stage generates smoothed NMF-TCSSBC ($x(n)$ where n is the frame index) from the speech signal using three steps. The first step computes magnitude spectrogram (V) of the speech signal. In the second step, the activation matrix H is computed from V using a dictionary W, which is learnt from the training speech spectra using NMF. The last step computes the NMF-TCSSBC from H matrix using equations 1 and 2 proposed by wang et al. [14]. This is followed by smoothing of the contour using a low-pass filter. In the second stage, peaks are detected from the smoothed NMF-TCSSBC contour using voicing decision. In second stage, peaks are also estimated using TCSSBC following the work on robust speech rate estimation (RSRE) [5]. The peaks from NMF-TCSSBC and RSRE are finally merged to compute the speech rate. The details of the dictionary learning and the main steps in two stages are discussed in the following sub-sections.

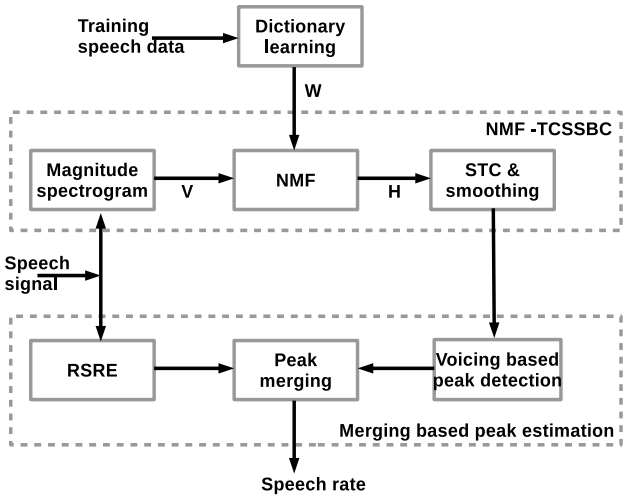

Fig. 2. Block diagram summarizing the steps of the proposed DB approach for speech rate estimation

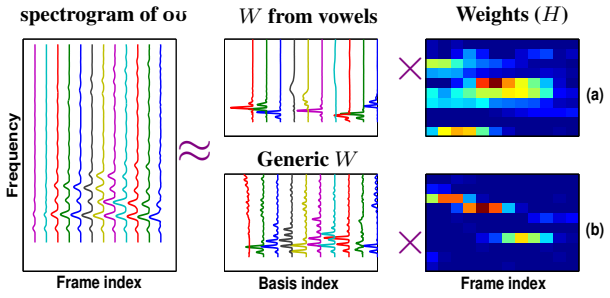

Fig. 3. Illustrative example to compare the dictionaries learnt from vowels and all speech frames.

3.1. Dictionary learning using NMF

The dictionary is learnt from the training speech spectra using NMF in which the non-negative magnitude spectrogram matrix V is factorized as the product of two low rank (r) non-negative matrices ($\approx W \times H$). The W is called a basis or dictionary matrix, that captures the structural bases of the training speech spectra. The H is called weight or activation matrix, that contains the activations of the projections of V on to the W matrix. The activations carry rich information about the spectra of segments, which are acoustically

similar to the ones used for training the dictionary W. This is illustrated with the help of Figure 3. The figure shows the factored matrices obtained from a spectrogram belonging to the vowel oʊ using two different W matrices – a) dictionary learnt from the vowel regions b) dictionary learnt from frames of all sound categories. In the Figure 3a, the H has consistent activations across all the frames and is less sparse, hence it results in high STC over all frames. In contrary to this, H in Figure 3b has less consistent and sparse pattern since a generic dictionary is used in this case. Though the W in the Figure 3 is learnt from the vowel regions, we experiment with different acoustic categories corresponding to broad phonetic classes and determine the best one experimentally. Once W is learnt, the magnitude spectrogram of the test utterance is factorized using NMF by keeping W fixed. The obtained H matrix used in the NMF-TCSSBC computation.

3.2. NMF-TCSSBC

For computing NMF-TCSSBC using H matrix, we follow the steps outlined by Wang et al. [14] for computing TCSSBC. We consider r rows of H matrix as r sub-band energies and compute the NMF-TCSSBC using the following steps – 1) apply the temporal correlation on r rows; 2) select M rows out of r rows with highest temporal correlation and 3) compute correlation across the M rows. However, the resultant NMF-TCSSBC can have noisy peaks, referred to as spurious peaks [5] [7]. We perform smoothning using Gaussian window of length L_s with variance (σ_s) to remove spurious peaks prior to peak detection.

3.3. Voicing based peak detection

Even after smoothing the NMF-TCSSBC contour, all peaks of the contour might not belong to syllable nuclei. The peaks belonging to syllable nuclei, called as syllabic peaks, are required to be detected effectively by discarding other peaks not belonging to syllable nuclei, called as non-syllabic peaks. In this work for detecting syllabic peaks in the NMF-TCSSBC contour, we use the peak detection strategy followed in RSRE method, consisting of the following steps – threshold (T_r) on the peak height with respect to neighboring largest minima; threshold (T_{dur}) on minimum duration between the two neighboring peaks; discarding the peaks belonging to the unvoiced regions. The detected number of peaks might not be equal to the number of syllables. The mismatch between the number of detected peaks and the number of syllables also happens with other existing techniques too [5] [7] [14]. However, the proposed DB method is found to be complementary to the RSRE method in the sense that it detects peaks which are not detected by the RSRE method. This is illustrated with the help of an exemplary sentence *Don't ask me to carry an oily rag like that* taken from the CTIMIT corpus shown in Figure 4. The figure shows the peaks detected by the proposed DB method (magenta) and the RSRE method (green). In the figure, the proposed DB approach detects extra peak at the syllable S_9 where the RSRE fails to detect. In such cases, merging the detected peaks of proposed DB approach with those from RSRE in an effective way could improve the speech rate estimation performance.

3.4. Peak merging

We propose a method to merge the syllabic peaks detected by RSRE and proposed DB method. Although both methods may detect one peak each within a syllable, the peak locations might not coincide. For example, in the Figure 4 the detected peaks belonging to the syllables S_1 – S_8, S_{10} – S_{12} do not fall at the same locations. Although

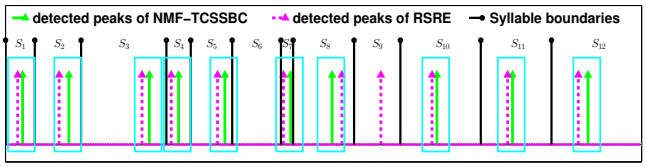

Fig. 4. Illustrative figure for explaining the complementary nature of the peaks detected from DB and RSRE methods. S_i denotes the i^{th} syllable.

Algorithm 1 Peak merging algorithm

1: **Inputs** window length: L_w, peak location sets: $\mathscr{X} = \{x_1, x_2, ..., x_N\}$ and $\mathscr{Y} = \{y_1, y_2, ..., y_M\}$,
2: Initialization: $\mathscr{V} = \mathscr{Y}$
3: **for** each location k in $\mathscr{Y}$ **do**
$k^* = \arg\min_i |x_i - y_k|$
4: **if** $(x_{k^*} - L_w) \leq y_k \leq (x_{k^*} + L_w)$ **then**
$\mathscr{V} = \mathscr{V} \backslash \{y_k\}$
5: **end if**
6: **end for**
7: $\mathscr{Z} = \mathscr{X} \cup \mathscr{V}$

their locations do not exactly match, they are closer compared to the peaks of two neighboring syllables. Such proximity in locations are indicated within cyan color boxes shown in the Figure 4. The peaks in each box are merged in the proposed peak merging algorithm. The steps involved in the peak merging are given in the Algorithm 1. The window length (L_w) used in the algorithm is determined experimentally. Following the steps in the algorithm, peak locations $\mathscr{Z}$ are obtained by merging the peaks from the RSRE method ($\mathscr{X}$) and the proposed DB method ($\mathscr{Y}$).

4. EXPERIMENTS AND RESULTS

4.1. Experimental setup

We consider the Pearson correlation coefficient (ρ) between the estimated syllable rate and the ground truth syllable rate for every sentence as the objective measure for evaluating the performance of the proposed DB approach. We consider RSRE technique as the baseline for the speech rate estimation. We learn NMF dictionaries for SWBD and TIMIT separately using training sets consisting of 100 audio segments selected randomly from each corpora. For the factorization, we use 'nmf_kl' function from the NMFlib v0.1.3 library [22] [23]. In the case of CTIMIT, we used the TIMIT dictionary, because CTIMIT is the noisy audio data re-recorded from the TIMIT[1]. For each corpus, 10 percent of randomly selected audio utterances are considered as the development sets for all three corpora. We learnt hyper-parameters – r, Gaussian window parameters (length L_t, variance σ_t) used in temporal correlation, L_s, σ_s, T_r, T_{dur} and L_w – separately for each corpus such that ρ values are maximized on the respective development set. Using these parameters, the performance is computed on test set containing entire corpus data excluding the training and development sets.

4.2. Hyper-parameter optimization

We experiment with learning dictionaries using frames from different categories of sounds. We choose regions belonging to vowels,

[1]Typically in the speech enhancement, NMF dictionaries are learnt from the clean audio and then later those are used for the enhancement of noisy data

Table 1. Optimal ρ values from the DB approach using the dictionaries leaned from different sound categories

	vowels	fricatives	stops	nasals
TIMIT	0.6137	0.5676	0.5779	0.5646
CTIMIT	0.3699	0.3529	0.3592	0.3628

Table 2. Hyper-parameters values optimized on the development set of all corpora.

	r	L_t	σ_t	M	L_s	σ_s	T_{dur}	T_r	L_w
TIMIT	15	3	1.6	5	5	1.5	15	25	5
CTIMIT	15	3	1.2	3	9	1.3	11	25	5
SWBD	15	3	1.6	7	7	1.5	9	25	9

fricatives, nasals and stops separately. Note that in ICSI SWBD, no phonetic transcription is available and hence, we do this experiment only on TIMIT and CTIMIT corpus. For SWBD, we use voiced regions for learning dictionaries. Using each dictionary ρ values are computed on the development set by optimizing the hyper-parameters. The maximum ρ values obtained for each case are shown in the Table 1 for TIMIT and CTIMIT. From the table it is clear that ρ is maximum for both TIMIT and CTIMIT, when the dictionary is learnt from the vowel regions. This is consistent with the findings from Jiahong et al [8] and Thilo et al [10]. They suggested that the acoustic properties of vowels correspond to the syllable nuclei hence vowel rate corresponds directly to syllable rate. The effectiveness of the vowel dictionary is also explained using an illustrative plot in the Figure 5 by taking an exemplary segment from the TIMIT corpus with the transcription '*Don't ask me to*'. Figure 5a and 5b indicate the H matrices obtained with vowel and stops dictionaries respectively. The figure suggests that using vowel dictionary captures the rich information about the vowels (/oʊ/, /æ/, /ɪ/, /ə/). Specifically activations corresponding to the vowel '/ɪ/, /ə/'are high in the Figure 5a than the Figure 5b. This indicates that the H obtained using vowel dictionary is beneficial in identifying the syllable nuclei. The optimal parameter set obtained with vowel dictionary for TIMIT and CTIMIT and voiced dictionary for SWBD are shown in Table 2. These parameters are used further to measure speech rate performance on the test set.

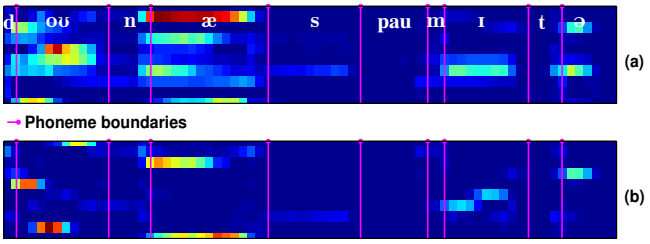

Fig. 5. Illustrative example for comparing H matrix by using dictionaries with rank (r) 10. The word corresponding to the illustrated portion is '*Don't ask me to*'.

4.3. Results and discussions

We compute the ρ on the test set for three corpora using RSRE as well as proposed DB method with merging (WM) and without merging (WOM) schemes. The ρ values are tabulated in the Table 3. In the case of CTIMIT, we denoised the audio files using a spectral subtraction technique with smoothing constants α and β as 0.98 and 0.6 respectively [24]. This is because the performance on the CTIMIT is found to be worse for all the schemes while the denoising improves the performance. The ρ values in the case of TIMIT and CTIMIT are lower using DB-WOM compared to those using RSRE. However, after merging the performance improves and the ρ values for TIMIT and CTIMIT using DB-WM are better than those using RSRE. This is because the peaks from the DB approach is often complementary to those from RSRE. This observation is illustrated with an exemplary sentence *Don't ask me to carry an oily rag like that* taken from the CTIMIT corpus in Figure 6. Figure 6a indicates peaks detected by the RSRE (cyan) from the smoothed TCSSBC. Figure 6b shows the peaks detected by the proposed DB method (magenta) from the smoothed NMF-TCSSBC. In both the figures syllable boundaries (black) along with syllabic transcriptions are indicated. The peaks at the syllables '/ɪ/ɪ/', '/l/ɪ/' are not detected by the RSRE. This could be due to smooth transition of the voiced consonant between neighboring vowels. However, the proposed DB method detects the syllable peak of '/l/ɪ/'. This is because the NMF-TCSSBC contour is more peaky than the TCSSBC at that syllable. We observe that total number of extra syllable nuclei detected by proposed DB approach is 6.52, 9.43 and 3.2 percent of total syllable nuclei present in the entire TIMIT, CTIMIT and SWBD respectively. No improvement in ρ for SWBD indicates that the complementary peaks detected by proposed DB method could result in more number of detected syllables causing a decrease in ρ. This could be because, lower vowel duration and high intra vowel acoustic variability in conversational speech [25] results less consistent structural information in the dictionary.

Table 3. Correlation coefficient on the test set using parameters optimized on the development set.

	RSRE	Proposed DB approach	
		(WOM)	(WM)
TIMIT	0.6590	0.6349	0.6794
CTIMIT	0.3466	0.2911	0.3646
SWBD	0.6550	0.6142	0.6413

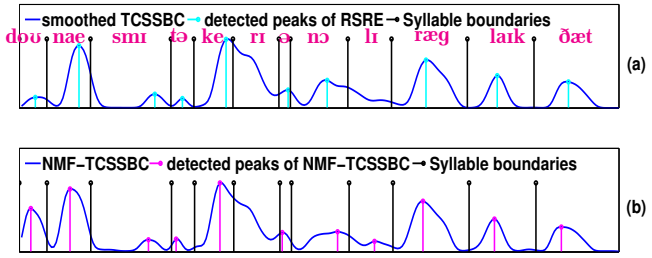

Fig. 6. Comparison between the peaks obtained from (a) RSRE method and (b) proposed DB approach.

5. CONCLUSIONS

We propose a dictionary based feature contour for the speech rate estimation task. The contour is computed from the weight matrix H obtained by NMF with a fixed dictionary W learnt from spectra of different sound categories. We find that the dictionary obtained from the vowel regions results in maximum benefit in speech rate estimation. Peaks representing the syllables are estimated by merging the peaks detected from the dictionary based contour with the peaks of traditional TCSSBC approach. Experiments with TIMIT and CTIMIT corpora reveal that the proposed DB method improves the performance compared to the TCSSBC method. Further investigation is required to develop a better peak merging strategy that could result in improvement for Switchboard corpus.

6. REFERENCES

[1] Nelson Morgan, Eric Fosler, and Nikki Mirghafori, "Speech recognition using on-line estimation of speaking rate.," *Proceedings Eurospeech*, vol. 4, pp. 2079–2082, 1997.

[2] Chris D Bartels and Jeff A Bilmes, "Use of syllable nuclei locations to improve ASR," *IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 335–340, 2007.

[3] Catia Cucchiarini, Helmer Strik, and Lou Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[4] Florian Hönig, Anton Batliner, and Elmar Nöth, "Automatic assessment of non-native prosody - annotation, modelling and evaluation," *International Symposium on Automatic Detection of Errors in Pronunciation Training*, pp. 21–30, 2012.

[5] Dagen Wang and Shrikanth S Narayanan, "Robust speech rate estimation for spontaneous speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2190–2201, 2007.

[6] Christian Heinrich and Florian Schiel, "Estimating speaking rate by means of rhythmicity parameters," *Proceedings Interspeech*, pp. 1873–1876, 2011.

[7] A Apoorv Reddy, Nivedita Chennupati, and B Yegnanarayana, "Syllable nuclei detection using perceptually significant features," *Proceedings Interspeech*, pp. 963–967, 2013.

[8] Jiahong Yuan and Mark Liberman, "Robust speaking rate estimation using broad phonetic class recognition," *IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 4222–4225, 2010.

[9] Tobias Cincarek, Rainer Gruhn, Christian Hacker, Elmar Nöth, and Satoshi Nakamura, "Automatic pronunciation scoring of words and sentences independent from the non-natives first language," *Computer Speech & Language*, vol. 23, no. 1, pp. 65–88, 2009.

[10] Thilo Pfau and Günther Ruske, "Estimating the speaking rate by vowel detection," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 945–948, 1998.

[11] Yaodong Zhang and James R Glass, "Speech rhythm guided syllable nuclei detection," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 3797–3800, 2009.

[12] Christian Landsiedel, Jens Edlund, Florian Eyben, Daniel Neiberg, and Björn Schuller, "Syllabification of conversational speech using bidirectional long-short-term memory neural networks," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5256–5259, 2011.

[13] Nivja H De Jong and Ton Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior research methods*, vol. 41, no. 2, pp. 385–390, 2009.

[14] Dagen Wang and Shrikanth Narayanan, "Speech rate estimation via temporal correlation and selected sub-band correlation," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 413–416, 2005.

[15] Tomas Dekens, Mike Demol, Werner Verhelst, and Piet Verhoeve, "A comparative study of speech rate estimation techniques," *Proceedings Interspeech*, pp. 510–513, 2007.

[16] JN Holmes, "The JSRU channel vocoder," *IEE Proceedings F (Communications, Radar and Signal Processing)*, vol. 127, no. 1, pp. 53–60, 1980.

[17] Daniel D Lee and H Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[18] Nasser Mohammadiha, *Speech Enhancement Using Nonnegative MatrixFactorization and Hidden Markov Models*, Ph.D. thesis, 2013.

[19] John J Godfrey, Edward C Holliman, and Jane McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 517–520, 1992.

[20] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at MIT: TIMIT and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[21] Kathy L Brown and E Bryan George, "CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 105–108, 1995.

[22] Graham Grindlay, "NMFLib - efficient matlab library implementing a number of common nmf variants," *URL-http://www.ee.columbia.edu/ grindlay/code.html*, 2010.

[23] Daniel D Lee and H Sebastian Seung, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, pp. 556–562, 2001.

[24] Yang Lu and Philipos C Loizou, "A geometric approach to spectral subtraction," *Speech communication*, vol. 50, no. 6, pp. 453–466, 2008.

[25] Francine Robina Chen, *Acoustic characteristics and intelligibility of clear and conversational speech at the segmental level*, Ph.D. thesis, Massachusetts Institute of Technology, 1980.