Published in final edited form as:
IEEE Trans Audio Speech Lang Processing. 2007 November 1; 15(8): 2190–2201. doi:10.1109/TASL.2007.905178.

Robust Speech Rate Estimation for Spontaneous Speech

Dagen Wang and
Viterbi School of Engineering, University of Southern California (USC), Los Angeles, CA 90007 USA. He is now with the IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA

Shrikanth S. Narayanan [Senior Member, IEEE]
Viterbi School of Engineering, University of Southern California (USC), Los Angeles, CA 90007 USA

Abstract

In this paper, we propose a direct method for speech rate estimation from acoustic features without requiring any automatic speech transcription. We compare various spectral and temporal signal analysis and smoothing strategies to better characterize the underlying syllable structure to derive speech rate. The proposed algorithm extends the methods of spectral subband correlation by including temporal correlation and the use of prominent spectral subbands for improving the signal correlation essential for syllable detection. Furthermore, to address some of the practical robustness issues in previously proposed methods, we introduce some novel components into the algorithm such as the use of pitch confidence for filtering spurious syllable envelope peaks, magnifying window for tackling neighboring syllable smearing, and relative peak measure thresholds for pseudo peak rejection. We also describe an automated approach for learning algorithm parameters from data, and find the optimal settings through Monte Carlo simulations and parameter sensitivity analysis. Final experimental evaluations are conducted based on a portion of the Switchboard corpus for which manual phonetic segmentation information, and published results for direct comparison are available. The results show a correlation coefficient of 0.745 with respect to the ground truth based on manual segmentation. This result is about a 17% improvement compared to the current best single estimator and a 11% improvement over the multiestimator evaluated on the same Switchboard database.

Keywords

Rich speech transcription; speech prosody; speech rate estimation; spontaneous speech processing

I. Introduction

SPEECH has been considered an attractive input modality for human–computer interactions for a long time. More recently, there has also been increasing interest in automatically mining vast amounts of speech data to determine not just what was spoken but how and by whom as well. Much of the research focus over the past three decades has been on automatic speech recognition, with tremendous progress being made, especially with the adoption of hidden Markov model (HMM)-based architectures. However, speech technology is still far from achieving the goal of robust speech understanding. One reason, which is also reflected in the current research trends in human language technologies, is the inability to adequately capture and represent the rich information contained in speech that is beyond mere speech-to-text transcription, as provided by conventional automatic speech recognizers. Humans use a wide

(wangdagen@yahoo.com; shri@sipi.usc.edu).

The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mark Hasegawa-Johnson.

variety of cues for recognizing and understanding speech, including intonation, prominence, and speaking rate. Machine processing of natural speech may also benefit from using these cues. Hence, one key goal of present day spoken language processing research is to automatically and robustly characterize these suprasegmental aspects of speech. This paper focuses on the topic of automatic speech rate estimation.

A. Significance

Speech rate is primarily dependent on two factors: speaking style and the nature/scenario of speech production (e.g., scripted/spontaneous). Research in this domain has two distinct application-driven threads as to how speech rate variability is addressed. On the one hand, variation in speech rate tends to adversely impact automatic speech recognition (ASR) and needs to be mitigated. On the other hand, variation in speech rate carries information critical for speech understanding and needs to be quantified to determine contextual variables such as speaking context, audience, knowledge of the subjects, etc. Much of the early focus on speech rate estimation was targeted toward improving ASR robustness. Even though HMMs have the ability to accommodate some of the spectral–temporal variations in speech, recognition accuracy is still severely influenced by mismatches between training and testing conditions. Speech rate variability is one such contributing factor [1]. A first step toward addressing this issue, i.e., to help improve the match between the models used and the speech being processed for recognition, is to quantify the inherent speech rate variability. Then, once an estimation of the underlying speech rate is done, one could select appropriately pretrained acoustic models [25], [54] or adaptively set transition probabilities of the HMMs [4], [5] that appropriately reflect the rate of the speech being measured.

Speech rate information can also be used in other speech processing scenarios besides robust automatic speech recognition. Speech rate variance could be interpreted as a function of the cognitive load associated with processing the text transcription [27], [42]. Cognitive load could be defined as the level of effort for the speaker/user to select the words to speak (for the main task or concurrent subtask [42]). In spontaneous speech scenarios, the speaker typically has to address various tasks on the fly, as they unfold, with unknown cognitive loads. So, not surprisingly, the speech rate variability for spontaneous speech can be quite large [44].

With increasing interest in spontaneous speech recognition and interpretation in recent years, and challenges posed by the acoustic and linguistic characteristics of spontaneous speech that are highly variable and more unstructured than prepared speech, the role of speech rate estimates has become ever more important. Notably, instead of just relying on the text from ASR to arrive at speech rate estimates, which may be quite noisy, there is a need to use suprasegmental acoustic features to directly facilitate speech interpretation. Below, we highlight some specific applications.

Prior research has shown that local speech rate correlates with discourse structure. For example, global analysis of the discourse structure in paragraphs and clauses has revealed that for each of the speakers considered, the average syllable duration of the first run of a paragraph is longer than the overall mean value per speaker in more than 60% of the cases (50% is the chance value)[3]. Local speech rate variations may carry other crucial information as well. For example, speech rate plays an important role in the context of sentence boundary detection and disfluency detection. It has been suggested that people tend to have longer syllable duration, or equivalently slower local speaking rate, at these events [6], [7]. Speech rate also correlates with prosodic prominence. Detection and normalization of rate of speech has been found to be necessary in measuring such attributes [8], [21]. Global speech rate also works as a normalization factor for many prosody-based classifiers. For example, it was selected as a key prosodic feature in the machine learning process of dialog act detection [19], [23]. In summary, speech rate estimation can be useful in a number of spoken language processing contexts.

B. General Measurement Methods

There have been two major trends in measuring speech rate. Each has its advantages and limitations. The first represents the use of discrete categorization—frequently, “fast,” “normal” and “slow”—to describe speech rate [24]. Such perceptually chosen classes have been used in applications such as acoustic model selection [9], [25] and HMM normalization [15] in ASR. Even though it matches human intuition, the boundaries between these three categories are fuzzy. Most of the time, human knowledge is required to set the boundaries, and hence it is difficult to devise a completely automated engineering solution.

In the second approach, speech rate is measured in a quantitative way by counting the number of phonetic elements per second. Words, syllables [9], stressed syllables, and phonemes [10] are all possible candidates, and syllables are a popular choice [6], [9], [11]. Studies on speech rhythm, i.e., organization of prominent and less prominent speech units in time, offer some motivation in this regards. Evidence from reiterative speech studies [16] supports the idea that syllable evolution is a good estimate of speech rhythm. Specifically, while the classic isochrony (or rhythm class) hypothesis regarding stress-timed, syllable-timed, or mora-timed languages has been largely unsupported by acoustic–phonetic evidence, a form of the isochrony hypothesis for rhythm has been shown to be supported by speech measures based on syllable structure and vowel reduction [50], [51]. Definitions for the syllable have been offered from a variety of perspectives; phonetically, Roach [37] describes a syllable as “consisting of a center which has little or no obstruction to airflow and which sounds comparatively loud; before and after that center (…) there will be greater obstruction to airflow and/or less loud sound.” This definition allows for a plausible way for detecting syllables in speech. Intuitively, syllables, by these accounts, should have an even distribution under normal speech production, and their rate could be changed as a result of speech rate change. Given such characteristics of syllables, the syllable-based rate estimate appears to be a widely used choice among speech rate researchers [6], [9], [11]. In this paper, we use number of syllables per second as a measure of speech rate. We will further explore the syllable’s acoustic property in Section II.

C. Role of ASR in Speech Rate Estimation

We first need to detect syllable boundaries for speech rate estimation. A straightforward, and convenient, approach would be through the use of automatic speech recognition where syllable boundaries can be retrieved as a side product of phonetic segmentation such as through Viterbi decoding [10]. Furthermore, ASR errors could be minimized with a supervised alignment process if the correct transcription were known [6], [7]. However, such an approach has limitations, while alternative approaches can offer other advantages.

First, assuming that the reference transcription is not available in real applications, recognition errors—especially for spontaneous speech—are unavoidable. Recognition errors (particularly insertions and deletions) would have the effect of degrading the performance of ASR-reliant speech rate estimation methods [25]. Second, speech rate could work as an acoustic feature to help ASR instead of being dependent on it. Hence, it is better to detect it in parallel or even be used as a part of an ASR front end. In this way, we can combine the complementary information produced by speech rate estimation and ASR. Finally, we believe that direct speech rate estimation can be easily extended to languages with vowel-centric syllable structures similar to English. This would be especially useful when only sparse data is available and where building a high-performance ASR system is especially challenging.

In this paper, we investigate using acoustic-only features to derive speech rate. The rest of the paper is organized as follows: Section II reviews the previous work and identifies the challenges. Section III introduces the data for evaluation. Section IV introduces our algorithm.

Section V describes the system and evaluation. The final section provides conclusion and discussion.

II. Previous Work

As stated in the previous section, we use number of syllables per second as the speech rate measure in this work. We therefore focus on identifying the correct number of syllables in an utterance.

A. Background

This task to identify the syllable structure in an utterance dates back to the very early stages of speech recognition research in the mid 1970s, where syllable detection was a popular first step in automatic speech recognition [41]. The HMM-based statistical framework for ASR had not been popularized then, and most of the research relied on knowledge (rule)-based acoustic signal processing. A variety of features had been proposed to capture the syllable nucleus. These included, for example, the use of linear predictive coding spectra [27], [28] or critical filter banks [32] to extract the low-to-high frequency energy ratios that characterize the acoustic properties of a syllable. Also, power spectra were used to derive a low-frequency profile in the region of first few formants of vowels [29], [30]. Due to restrictions of processing hardware and data availability at that time, these efforts were limited to read speech in quiet laboratory environments, usually produced as isolated words or slow, carefully read sentences. [41].

With the wide adoption of hidden Markov model-based speech recognition in the 1980s, there was a decreased focus on acoustic–phonetic studies for ASR. However, recently with the increased scope of spoken language processing (Section I) the need for processing meta-linguistic features has increased considerably, resulting in many interesting approaches, including for speech rate estimation [8], [12]. A significant advantage of present-day research is the ability to use large, spontaneous speech corpora to obtain statistically significant results. An influential recent effort on speech rate estimation is by Morgan and Fosler-Lussier [9]. Our paper was inspired, and builds upon on their work, which we will review further in Section II-C.

All of the previously proposed techniques share the basic knowledge-based feature extraction ideas. The strategy relies on converting the speech waveform to a lower (frequently, one)-dimensional representation. Following that step, the syllable nucleus is located by picking peak patterns in such a representation. There are alternatives to simple peak picking. For example, Mermelstein [29] used a “convex hull” algorithm to recursively detect peaks which are prominent relative to their surroundings. Rabiner used a static threshold on the total energy profile [31].

In addition to these rule-based approaches, there have been attempts to use statistical learning methods to derive syllable nuclei. Normally, a large number of features are extracted such as log energy spectra organized in critical bands [33], bark scale filter bank [34], and even auditory models (RASTA [35]). The learning methods are mostly based on hidden Markov models [34] or artificial neural networks [33] and are usually trained with appropriately annotated corpora.

B. Acoustic Characteristics of Syllables

The task of automatically detecting the syllable nucleus has a close relationship with vowel landmark detection [41] based on the assumption that a syllable is typically vowel centric and neighboring vowels are always separated by consonants. The use of the term “vowels” in this context can be in fact generalized to “sonorant segments,” in light of the discussion in Section I-B about the definition of syllable. Generally speaking, vowels form the nucleus of syllables, whereas consonants form the boundaries in between [40]. However, it should be noted that a more precise characterization of the syllable structure can be made in terms of sonority (a sound's "loudness relative to that of other sounds with the same length, stress, and pitch." [40]) which posits that syllables contain peaks of sonority that constitute their nuclei and may be surrounded by less sonorous sounds [52], [53]. According to the Sonority Sequencing Principle [52], vowels and consonant sounds span a sonority continuum with vowels being the most sonorous and obstruents being the least, with glides, liquids, and nasals in the middle. In this paper, we will use the term vowels to mean sonorant sounds in the nucleus of a syllable. We use this convention for simplicity and because vowels constitute the most sonorous and frequent members of syllable nuclei.

A vowel is characterized by an open configuration of the vocal tract so that, unlike consonants, there is no significant build-up of air pressure above the glottis [40]. Due to resonances in the vocal tract, a vowel exhibits clear formant structure in its spectrum. This contrasts with consonants, which are characterized by a constriction or closure at one or more locations along the vocal tract. We will use this general description to motivate our design of the algorithm for syllable nucleus detection.

C. Subband-Based Correlation Approach

As a preface to the description of our algorithm, we review the correlation based approach proposed by Morgan and Fosler-Lussier [9] and other related work. One classic way to get syllable counts is through performing full-band spectrum/energy analysis and measuring the dominant peak of the long-term envelope [13]. However, such an approach results in significant noise in the final envelope, making it difficult to obtain syllable counts robustly.

Many further improvements for the energy/spectrum idea have been proposed. For example, Pfitzinger [20] extracted a band-pass signal and applied rectifying and smoothing window to it before performing peak counting. In that work, a 21.8% error rate (a measure that uses syllable nucleus matching between test and transcription location) was reported. As an alternate approach to the same problem, the first spectral moment of the broadband energy envelope was used as a speech rate measure [12]. While this method provided improved performance with conversational speech, it was shown that using a one-hour subset of manually transcribed Switchboard data, the correlation between transcribed syllable rate and the experimental rate was only about 0.4 (when both were measured over between-pause spurts) [12].

All the aforementioned syllable detection approaches assume the rate of peaks on wide band energy envelope (see, e.g., Fig. 1) is a valid representation for speech rate measure. However, this assumption has its limitations. For instance, formant structure, which is crucial for syllable identification in fast speech, is lost when the wide band energy envelope representation is used. For example, the same magnitude on wide band energy envelope might correspond to different formant structure, thus different vowels. For fast speech, the transition between different vowels is difficult to identify by energy envelope. Since such a wide band energy envelope is only one of many possible representations of speech, researchers have proposed alternative measures. One of the major improvements was given in [9], where Morgan and Fosler-Lussier developed a subband-based module that computes a trajectory that is the average product over all pairs of compressed subband energy trajectories. That is, if $\boldsymbol{x_i(n)}$ is the compressed energy envelope of the $\boldsymbol{i}$th spectral band, a new trajectory $\boldsymbol{y(n)}$ is defined as

$$y(n)=\frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}x_i(n)\,x_j(n) \tag{1}$$

where N is the number of bands, and $M = N(N - 1)/2$ is the number of unique pairs. The algorithm and the system of [9] is summarized in Fig. 2. By this method alone, correlation coefficients greater than 0.6 were achieved between the referenced and measured speech rate values. Furthermore, it was shown in [9] that the performance would boost to 0.673 if multiple estimators were combined (with wideband energy peak count and spectral moment count; see Fig. 2). This method addresses the formant structure issues we discussed earlier by introducing band-wise correlation in the spectral domain, which accentuates the syllable peak in the correlation profile.

We build upon this method, and address two key challenges. The first one relates to choosing the robust feature set to identify the syllable nucleus. Solutions have been proposed from both signal processing [27], [29] and speech production [35] points of view. We consider both spectral and temporal features in characterizing the syllable envelope as described in Section IV. The second problem concerns optimal parameter selection. Heuristic methods have been popular, but they do not guarantee optimality or generalizability across domains [31]. Statistical learning schemes are attractive in the sense of objectively trying to seek optimal parameters. The challenges, however, include the availability of an appropriate training scheme, and effectively dealing with multiscale, multidimensional features such as those needed for the speech rate problem [34]. We adopt a Monte Carlo simulation-based method, followed by a systematic sensitivity analysis to facilitate parameter estimation. We evaluate our method on a database of spontaneous speech, which we describe in Section III.

III. Database

Our primary goal is to robustly detect speech rate on spontaneous speech. We use the phonetically transcribed ICSI Switchboard corpus subset (provided kindly by Fosler-Lussier [9]). Switchboard is a corpus of several hundred informal speech dialogs recorded over the telephone [11], [39]. The corpus is extensively used for development and testing of speech recognition algorithms and is considered to be fairly representative of spontaneous discourse. In contrast to carefully enunciated, read speech (such as TIMIT [43]), the speech contained in Switchboard tends to vary significantly in terms of rate, prominence, etc. A total of 5682 spurts were hand transcribed phonetically by linguists in the Switchboard Transcription Project at ICSI [2]. The transcription includes syllable boundary information (not manually segmented but hand-corrected machine derived segmentations). The cutoff marks (h#, sil) are taken care of to get the accurate reference syllable numbers. This corpus is the same as used in [9].

IV. Algorthm Design

Our proposed algorithm works by abstracting the speech waveform to a 1-D envelope and detecting syllables by peak picking. It consists of four stages: spectral processing, temporal processing, smoothing, and thresholding. This section is organized in the following way: First, we will summarize a number of practical issues that the algorithm needs to tackle. Second (in Sections IV-B–E), we will describe the four stages of our algorithm, clarifying which particular challenge each part is addressing. Finally, we will describe our strategy for choosing the optimal parameters for each algorithm setting.

A. Practical Challenges

Our algorithm is based on the speech subband correlation approach [9]. Peak picking on the resulting correlation envelope gives the syllable number estimation. A major challenge is due to noise in this envelope, that can result from a variety of sources as discussed below, and can interfere with the peak picking and degrade the accuracy of syllable number estimation.

1) Background Noise—Background noise is a significant contributing factor toward spurious peaks in the correlation envelope. For example, in Fig. 1, there are instances of background noise in the regions between 0.78 and 0.85 and 1.05 and 1.15 s. Such noises tend to introduce extra peaks in the final correlation envelope. One traditional way is do noise cancellation or suppression. However, often, noise can be of disparate types, and difficult to characterize. Such noises also include soft breath and cross-channel voices that are not a part of the foreground speech. We apply pitch verification and relative thresholding techniques to address these problems.

2) Consonant “Noise”—Consonants are key components of speech. The particular correlation approach we consider here, however, relies on vowels to be the major contributor of the syllables and thus the peaks. As explained in Section II-B, additionally this includes sonorant consonants, such as /l/, /r/, which can also carry syllabic weights. However, other (obstruent) consonants, especially fricatives, also sometimes contribute extra peaks not related to the “syllable peak.” This is why they are categorized as “noise” here. The characteristic of such noise is that they do not have as much energy as a vowel. Furthermore, they may not have pitch associated with them when they are unvoiced. Lastly, they normally have short durations. We will show how these cues can be advantageously exploited to mitigate the effects of consonant “noise.”

3) Smearing—In our experiments, and also those in [9], there are a number of individual cases where a high speaking rate sometimes results in smearing neighboring energy peaks. This makes it particularly difficult to derive the correct number of syllables for that segment.

Fig. 3 shows an example of smearing of syllables “in” and “tro” (from the word “introduction”) showing only one peak. The possible reasons include effects of windowing used in the analysis and any smoothing of the envelope in a post processing step.

4) Overestimation Issues—It is also observed that for some slow segments, people tend to shift the vowel formant to express some prosodic content. Such phenomena will bring extra peak estimates in the direct application of the subband correlation method as proposed in [9].

In the example shown in Fig. 4, “so” has only one syllable. With a fixed subband, when a formant shifts from one band to another, it will generate an additional peak.

5) Windowing Effect—In all these methods, a correlation envelope is generated and utilized. Like all short-time windowing methods, a larger window makes the envelope smoother but loses fine details. A smaller window provides more detail but makes the envelope noisy and in turn renders peak counting difficult. In the syllable scale we are considering, such windowing effects are unavoidable. We will address this problem by Gaussian filtering.

The aforementioned challenges are addressed in the four steps of the proposed method, as described below: spectral processing, temporal processing, overall smoothing, and thresholding.

The overall system flow chart is shown in Fig. 9.

B. Spectral Processing

1) Selected Subband Correlation—We believe formant structure is the major key to identifying vowels and thus locating the syllable nucleus. Our algorithm aims to abstract the speech waveform to a 1-D envelope, with a general strategy to let the center of the vowel to be maximized while not significantly increasing the contribution to the envelope from the consonants. As a consequence, the neighboring syllables (vowel centric) should have a deeper gap in between. The subband correlation addresses this issue. We wish to further improve its performance by doing a selected subband correlation.

In all previous approaches, spectral correlation is performed on the full bands. However, we find that if we concentrate on the prominent subbands where the formant structure lies, the vowel segments could be further boosted while the consonant contribution will be diminished comparatively. Such discrimination increase will be useful for later threshold setting. So we propose to do spectral correlation only on a selected subset of the subbands. First, instead of choosing only four subbands, we apply a 19 subband analysis (by a facility provided in the speech filing system tool [14]). We then keep the top *M* bands by subband energy for further temporal and spectral processing. *M* is a parameter we need to set appropriately and will be discussed in a later section.

In the example shown in Fig. 4, slow speech incurs an overestimation of syllable number, and we noticed that the formant structure has shifted within the vowel segment. In this case, if we select the top *M* most prominent subbands to do correlation, the shifting effects could be automatically tracked and resolved.

2) Pitch Verification—In the previous section (Section IV-A), we outlined the characteristics of background noise and consonant noise. Typically, such regions do not have any voicing. The availability of pitch information could serve to identify this effect. Pitch estimation is a fairly mature signal processing technique and can be easily implemented using a variety of approaches. The use of pitch in conjunction with the correlation envelope could help eliminate the pseudopeaks where there is no pitch. In this paper, we apply the pitch estimation algorithm that is based on normalized cross correlation function and dynamic programming. It is similar to that as presented in [46]. Such an approach was found to be very effective as shown in the later evaluation section.

C. Temporal Processing

Few previous approaches incorporate temporal processing. However, we note that each landmark lasts over some period of time. For example, vowels and sonorant consonants which constitute the major body of a syllable extend over several tens of milliseconds. Silence and nonsonorant consonant sounds can also cause signal discontinuity in the temporal realm (consonant discontinuities are typically shorter). Temporal processing, aimed at achieving desirable smoothing effects, is carried out as described below.

1) Temporal Correlation—Inspired by spectral cross correlation, and also by the fact that each syllable (i.e., similar spectral pattern) typically lasts over several tens of milliseconds, we also perform a cross correlation in time domain.

Let $x_t, x_{t+1}, \dots, x_{t+K-1}$ represent an increasing time order of subband energy vectors with length *K*. We then compute correlation y_t as

$$y_t = \sqrt{\frac{1}{2K(K-1)} \sum_{j=0}^{K-2} \sum_{p=j+1}^{K-1} x_{t+j} \bullet x_{t+p}}. \tag{2}$$

Through this correlation, each syllable has a peak at its center, because it spans most of the part of this syllable.

It also could be viewed as a type of filtering. However, compared to linear weighting of neighboring frames, the above approach uses products which will boost within-syllable frame similarities. This approach was found to effectively address the windowing effect of the envelope. The parameter we need to set here is *K*, the size of the window to do the correlation.

2) Weighting Window—Fig. 3 illustrates the case where fast speech has a smearing effect on neighboring syllables. Even though the major purpose of our algorithm is to smooth the correlation envelope, we do not want to lose important details in the process. In order to emphasize intersyllable discontinuities, we apply a Gaussian weighting window centered at the middle of the analysis frame before the process of self temporal correlation (as described in Section IV-CI). So the center part, in the case there is a small discontinuity, is amplified, and this frame has more weight in the correlation process. Such an approach could be mathematically described as follows.

Let w_0, w_1, … ,w_{K-1} represent a series of window coefficients. We first perform a weighting operation on the subband energy temporal vector series x_0, x_1, … ,x_{k-1}

$$x_{t+j}=w_j x_{t+j}. \tag{3}$$

Here, we choose *w* to be a Gaussian window centered in the middle of the analysis segment. After this process, we plug in the updated vector series x_0, x_1, … ,x_{k-1} to the temporal correlation process as described in Section IV-C1. We need to set the variance of the Gaussian window appropriately to control the shape of the window.

In order to illustrate the effects of such weighting window, we study the discontinuities of the step function in the 1-D case, and show the results in Fig. 5. (The temporal correlation in Section IV-C1 is for *M*-dimensional vectors where *M* is the number of selected subbands.)

The original step signal has the sharpest edge. The effect of the weighted windowing, as can be seen in Fig. 5, is to help reach an acceptable tradeoff between amplifying the discontinuity while achieving the desirable smoothing effect suitable for rate detection. These parameters used in correlation and weighting the window are selected as optimal settings for the experiments in Section V where we will further discuss the implications of this algorithm.

It also needs to be mentioned that there are many possible filter selections to achieve similar smoothing effects. Gaussian windows offer desirable kernel characteristics and easy parametric control of their shape, and are widely popular in image processing for smoothing [47]. Also, both the Fourier transform and the derivative of a Gaussian window are Gaussian functions. We hence adopt the 1-D Gaussian window for our case.

D. Smoothing

After the spectral and temporal correlation, we obtain a 1-D correlation envelope. There still may be local peaks in the correlation envelope which result in spurious peak counts. As a result, some type of further smoothing becomes necessary. We apply the standard Gaussian filtering method. The parameter setting strategy for the filter is described in Section IV-F.

It needs to be clarified that our algorithm has two different Gaussian windows involved with different intended use. While the purpose of the one described in Section IV-C is to alleviate the smoothing effects of the temporal correlation by making the slope sharper, the purpose of the one in Section IV-D is purely to provide a low-pass smoothing filter.

E. Threshold Mechanism

In addition to smoothing, for handling spurious peaks in the correlation envelope, we could design further thresholding mechanisms to improve the overall robustness of the peak counting.

Based on empirical analysis on several speech correlation envelopes, we categorized the observed spurious peaks into 2 two classes: First are those that occur when there is no voicing activity. We proposed in Section IV-BII to use pitch verification as a hard threshold where all peaks with no corresponding pitch activity are removed. However, there are limitations to pitch verification such as when there are voiced consonants, cross-channel voice, or pitch computation error. The major characteristic of such noisy peaks is that they are of relatively low amplitude. Such peaks could be removed by appropriate thresholding. The second class of noisy peaks appears in the voiced part. In this case, neither pitch verification nor absolute thresholds would be effective since those regions always have nonzero pitch, and the noisy peaks are of quite high amplitude. Most algorithms in Sections IV-A–D try to address this issue to some extent. As an additional step, we design a threshold mechanism which could deal with pseudovoiced peaks specifically.

1) Temporal and Magnitude Thresholds—To counter pseudopeaks that occur close in time, first, we set a threshold for the minimum distance in time between two neighboring peaks. The simple idea here is that two syllables could not be very close in the final correlation envelope with respect to the frame advance of 10 ms. Second, we still need to set thresholds on magnitude.

Fig. 6 illustrates a case where a single syllable displays two peaks (marked peak A and peak B) in the final correlation envelope. We propose to measure the minimum difference between a local peak and its larger neighboring minima instead of the ground zero, for setting temporal thresholds. For example, in Fig. 6, the threshold magnitude of peak A is measured by the relative magnitude between A and C; similarly, for peak B it is measured between B and C. This method however could fail to report any peaks in specific cases such as in Fig. 6 (Since the relative magnitudes of peak A and peak B are all very small). Instead, we found that a modification that considers the magnitude of a peak with respect to its immediate preceding minimum to be more robust. This was based on observations about typical syllable-level acoustic characteristics that demonstrate larger ranges between neighboring syllables, i.e., high absolute magnitude (such as A or B) at the syllable and rather low absolute magnitude between the neighboring syllables (such as D, E). On the other hand, spurious peaks tend to have smaller ranges. Hence, in the new scheme, for example in Fig. 6, peak A’s threshold magnitude is measured by the relative magnitude difference between A and D. Peak B’s threshold magnitude is measured by the relative magnitude difference between B and C. Peak A could thus pass the threshold since it is rather high in such magnitude. So, it returns the correct peak number.

This scheme could also handle many other cases very well. In the case that A–D and B–C are very close and high, this most probably implies that they are two distinct syllables and the algorithm will keep both. If A–D and B–C are both of small magnitudes, considering D has low absolute magnitude, they are both removed as background noise. The other advantage is that this left-compare-only threshold is compatible with absolute thresholding: When we apply it on silence regions, this method works the same as absolute threshold.

It should be noted that there is potential failure possibility of this threshold mechanism in the case of very close syllables with no discernable boundaries such as in the words “reenter,” and “reenergize” which may appear as pseudovoiced peaks in fast speech. Nevertheless, overall, we expect that these cases to be relatively infrequent, and that the proposed threshold mechanism would be in general effective.

F. Parameter Selection

The previous sections described many approaches for improving the syllable detection performance robustness. One critical question that still needs to be answered is how to choose the different algorithm parameters to enable the various processing blocks to work well together. The manual heuristic method has its own merits in that it utilizes expert human knowledge for rapid parameter setting. This is especially useful when a single running cycle (even on development test set) is computationally intensive. However, the approach suffers from limitations of scalability. For instance, many iterations of tuning may be needed, and it may be difficult to tell when the algorithm reaches a local maximum or if we could find the global maximum. Furthermore, such an approach would be difficult to easily port to other data types and domains. Hence, we propose to use a principled way for parameter estimation relying on Monte Carlo-based initialization followed by a sensitivity analysis to set the parameters using a development set.

1) Monte Carlo Method—The algorithm we have proposed for speech rate estimation poses a multidimensional parameter setting problem. We adopt the Monte Carlo method to bootstrap the parameter value initialization. The first step is generating the possible ranges for the parameter values. We specify these initial ranges rather large (greedily) and then generate the parameter set by Monte Carlo sampling. Fig. 7 illustrates the sample histogram after 4446 runs on the development set. The algorithm's performance with the selected parameters is then noted. The large initial parameter set requires that a large number of random parameter samples be generated in order to reach the optimal region, a computationally intensive process. We made this possible by optimizing the batch operation and offline front-end processing. Since Monte Carlo simulation draws parameters randomly within a large range, it is an important step towards detecting the global maxima. Though with a given number of simulations, we cannot guarantee to find the global maxima, we believe it at least provides an acceptable approximation to it.

2) Sensitivity Analysis—The chosen parameter values were then subjected to a sensitivity analysis. This was done through systematic perturbations to the parameter values (obtained from the Monte Carlo simulation) until a local maximum is reached. We first define an "atomic increment," which specifies the smallest amount by which each parameter could change. We then perturb each parameter one by one with the atomic increment in each direction. Every time there is an improvement, we will update the relative parameter. This step is repeated until no further improvement is obtained for perturbations on all parameters.

In Fig. 8, the X -axis shows the number of the perturbation trials. This number starts from 0 and increases by the aforementioned procedure. The Y -axis shows the correlation coefficient between speech rate estimates obtained from the test and reference data in the development set. The correlation coefficient is an indicator of speech rate estimation accuracy. Fig. 7 then illustrates how such perturbations could monotonically improve the performance. We found for fast convergence, the Monte Carlo method is essential to obtain a good rough estimate of the starting point. The sensitivity analysis is designed in such a way to efficiently but exhaustively search the parameter space to scan all possible local maxima in the given range.

V. System Description and Experimental Results

Given the description of the various components of our algorithm in Section IV, we will now describe the full system and report the evaluation results.

The overall speech rate estimation system is summarized in Fig. 9. Each block therein was described in Section IV. The algorithm parameters are set systematically and automatically using the Monte Carlo simulation and sensitivity analysis described in the previous section.

The technical specification of each functional component is described below in order.

- The speech is passed through a 19-channel filter bank analyzer to get the energy vector series. We apply the utility "voc19" as provided by [14]. It is a straightforward implementation of a 19-channel filterbank analyzer using two second-order section Butterworth bandpass filters spaced as in [22]. Energy smoothing is done at 50 Hz to give a default 100-Hz frame rate. Here we do not apply any energy compression procedures as in [9].

- With such a 19-channel filter bank, we get a 19 stream subband energy series. Only the top bands are selected and kept.
- Then we choose *K* temporal frames. These *K* frames are weighted by a Gaussian Window as described in Section IV-C2. Temporal correlation is then applied as detailed in Section IV-C1. The overlap across successive Gaussian windows is K−1 frames.
- For the next step, the resulting subband energy vector is cross-correlated in a way identical to [9].
- Finally, peak counting is performed on the final smoothed envelope with pitch validation and various thresholding schemes as in Section IV.

In order to set the parameters, we randomly selected 568 speech spurts from the full ICSI Switchboard data set as the development set which represents about 10% of the data. Applying the Monte Carlo simulation and sensitivity analysis, we obtained the parameter values as listed in Table I.

While this is a multiparameter tuning problem, it is also desirable to understand the effect of the individual parameters. To experimentally obtain insights in this regards, we evaluated the performance by removing each of the proposed component and measured the resulting performance on the development test set. Following methods in [9], a transcribed syllable rate was computed by dividing the number of syllables occurring in the spurts by the length of the spurt. In this paper, we treat this rate as the reference rate. We use the detected rate to correlate with the reference rate to get the final agreement measure on the data set. We also computed the simple mean squared error (MSE) between the estimated and reference rates as follows:

$$\text{MSE}\% = \frac{\text{ErrorEnergy}}{\text{referenceEnergy}}^{*} 100\% = \frac{\|\ \text{estRate} - \text{refRate}\ \|^{2}}{\|\ \text{refRate}\ \|^{2}}^{*} 100\%.$$

The results are reported in Table II.

All the components appear to provide improvements in the performance, but to varying degrees: Results show pitch validation to be the most effective, with thresholding strategies also contributing significantly on this data set. The use of reduced, instead of full, number of bands improves the error variance without degrading (in fact, slightly improving) the correlation rate and MSE, but with obvious reduced computation.

While interpreting the results of Table II, we should note that the algorithm was designed to have several mutually dependent components working together to locate the syllable nucleus correctly. As motivated in Section IV, each component attempts to address specific issues in rate estimation, and the Monte Carlo approach enabled us determine a compromise optimum of these parameters. Hence, the method of evaluating relative performances by turning off components with respect to a jointly tuned parameter set may not necessarily assure optimal settings for the remaining components. The only exception to this might be the pitch validation component. Since the computation of pitch is independent of all other components, its contribution is most likely also largely independent of the other modules. Table II shows that the performance degradation by turning this option off is the most significant. This implies that it could remove the effects of background and consonant noise (Section IV-B2) which are difficult to be mitigated by other components. The results also suggest that the pseudopeaks removed by pitch validation constitute a significant portion of the impediment to accurate rate estimation.

The results of Table II also indicate that the thresholding schemes contribute noticeably to the system performance. However, the contributions do not come just from the “threshold” selection but from the effects of other signal processing components that help isolate the “noise” that is then easily removable by thresholding. For example, subband correlation helps to boost the contribution of vowel and other sonorants while suppressing the intersyllable valleys. This makes the margins between true peaks and pseudopeaks accentuated, which in turn facilitates the thresholding schemes to work robustly.

Temporal correlation and Gaussian filtering both try to achieve the same goal of smoothing the syllable envelope. Table II shows that they contribute similarly to the overall system. We believe that the joint parameter setting with Monte Carlo approach would set these two subsystems to work optimally with the thresholding scheme. In sum, the experiment of studying the effects of the various components shows their relative importance, although it is understood their settings in this process may not be entirely optimal.

In the next step, we proceeded with the evaluation of the full system with all the available Switchboard data and the parameter settings obtained from the Monte Carlo simulation and sensitivity analysis, again following the methods as reported in [9]. We use the detected rate to correlate with the reference rate to get the final agreement measure on the full 5682 spurts set. Also, the mean error and standard deviation error were calculated. The results are reported in Table III. This result represents about 17% improvement compared to a single estimator and 11% improvement with respect to a multiestimator evaluated on the same database in [9].

Also, instead of using all of the switchboard data and removing just the development part, the correlation coefficient is 0.734, which is slightly lower than the results in Table III.

In addition, we analyzed the influence of certain factors on the estimation of speech rate. In Section II-B, we noted that besides vowels, sonorant segments of syllable nuclei might include glides, liquids, and nasals. In Table IV, we report results for the two cases separately: speech spurts which have at least one syllable with glides/liquids/nasals as the sonorant elements, while the other class consists of spurts with only vowels as syllable nucleus. Results show that the inclusion of sonorant consonants is handled well by the algorithm.

We also investigated the effect of the actual value of the speech rate itself. For that purpose, we heuristically categorized the speech data into three classes based on transcribed speech rate: fast (> 5 syllables/s, 711 spurts), normal (between 3 and 5 syllables/s, 3405 spurts), and slow (< 3 syllables/s, 1566 spurts). The estimated and reference values are shown in Fig. 10 for each of these data conditions. In general, the estimated values tend to be underestimates, with greater disagreements in the case of slow and fast speech (second and fourth panels in Fig. 10). We calculated the mean squared error between the reference and estimated values for each of these cases: the overall MSE rate was 5%, while the rates for slow, normal, and fast cases were 10.3%, 3.5%, and 6.8%, respectively. The major cause of this effect is due to two factors: overestimation and smearing, which occur often in slow and fast speech, respectively. (Refer to Sections IV-A3 and A4).

It needs to be clarified that the number of syllables per utterance might be an ill-defined quantity. Even though we use the normalized syllables per second as the rate measure, this quantity might not keep constant as the spurts length is varied. This should be taken into consideration for the justification in Fig. 10.

Lastly, we wish to explore yet another property of our algorithm. Throughout this work, we have been assuming that the peak number is a valid indication of the syllable number. It assumes that the peak location on the correlation envelope should be consistent with the syllable location. Even though this is not part of the work of Fosler-Lussier and Morgan [9], and it might not be a necessary condition to make our algorithm work, we include these statistics for closer analysis. For this purpose, we treat the original syllable transcription in the ICSI Switchboard corpus subset as a “gold standard.” Then, we compare the peak location on the correlation envelope to this standard. If within a syllable, there is a one-to-one mapping, we treat this as “correct.” Otherwise, it is deemed as a “deletion” or “insertion.” The statistics are provided in Table V.

For a spontaneous speech corpus like Switchboard, more than 80% of the time the syllable gives a one-to-one mapping. As stated in Section IV-A, our algorithm has deletion and insertion errors under specific circumstances. Even though slow speech rate is slightly more difficult to estimate (as illustrated in Fig. 10), due to the preponderance of the number of fast-spoken syllables relative to the slow-spoken ones in the data, deletion errors dominate insertion errors.

It should be noted our algorithm is optimized towards improving the speech rate correlation between a reference and the measured, and it might not necessarily produce the optimal syllable location information. One reason, as discussed in Fig. 10, is the ill-defined nature of syllables per utterance as a rate indicator.

VI. Summary and Conclusion

Our experiments show that the speech rate estimation methods proposed in this paper offer further improvements over previous methods. Such advantages are demonstrated by improved correlation coefficients and reduced mean error and standard deviation in the estimates with respect to the reference values. We have also systematized the heuristic parameter setting methodology originally used in [18]. The Monte Carlo method and dynamic parameter perturbation schemes provide ways for parameter tuning that guarantee finding the local maximum and approximating the global maximum. For the Monte Carlo method, the coverage is large, but the precision is low. Local convergence is achieved in postprocessing through sensitivity analysis implemented through systematic parameter perturbations. Such a dynamic perturbation scheme could help find the neighboring local maxima but cannot guarantee to enumerate all the local maxima.

The key part of the algorithm is in obtaining the correlation envelope. Such a signal envelope measure could disclose other useful information like syllable duration and spectrum intensity. For example, in [45], this envelope was used to derive a measure for word prominence.

There are further avenues that can be considered for improving the methods presented in this paper. For instance, it is well known that there are a number of factors that could affect the phonetic characteristics of a syllable (duration, f0), notably the underlying linguistic prosodic structure, which can impact the syllable detection accuracy, a critical aspect of the speech rate measure proposed in this paper. Specifically, lengthening at the edges of prosodic domains (boundaries) has been well documented both in read speech [49] as well as in spontaneous speech [48]. This includes the effect of utterance position: initial words are longer than noninitial words; utterance final words are longer than utterance medial words. These in turn can influence the quality of automatic syllable detection that relies on the acoustic characteristics of the syllable. Explicitly incorporating contextual information, such as the temporal structure, can further improve the proposed algorithm.

A possible alternative would be designing an adaptive algorithm for dynamic parameter adjustment such as through multipass rate estimation. For example, the first pass can give a rough estimate of the rate, while the second pass can use the results of the first pass to set relative parameters. Such an approach could be implemented iteratively. However, in applications of rate estimation that require real time processing, such multiple-pass methods may drastically limit the usefulness of rate estimation.

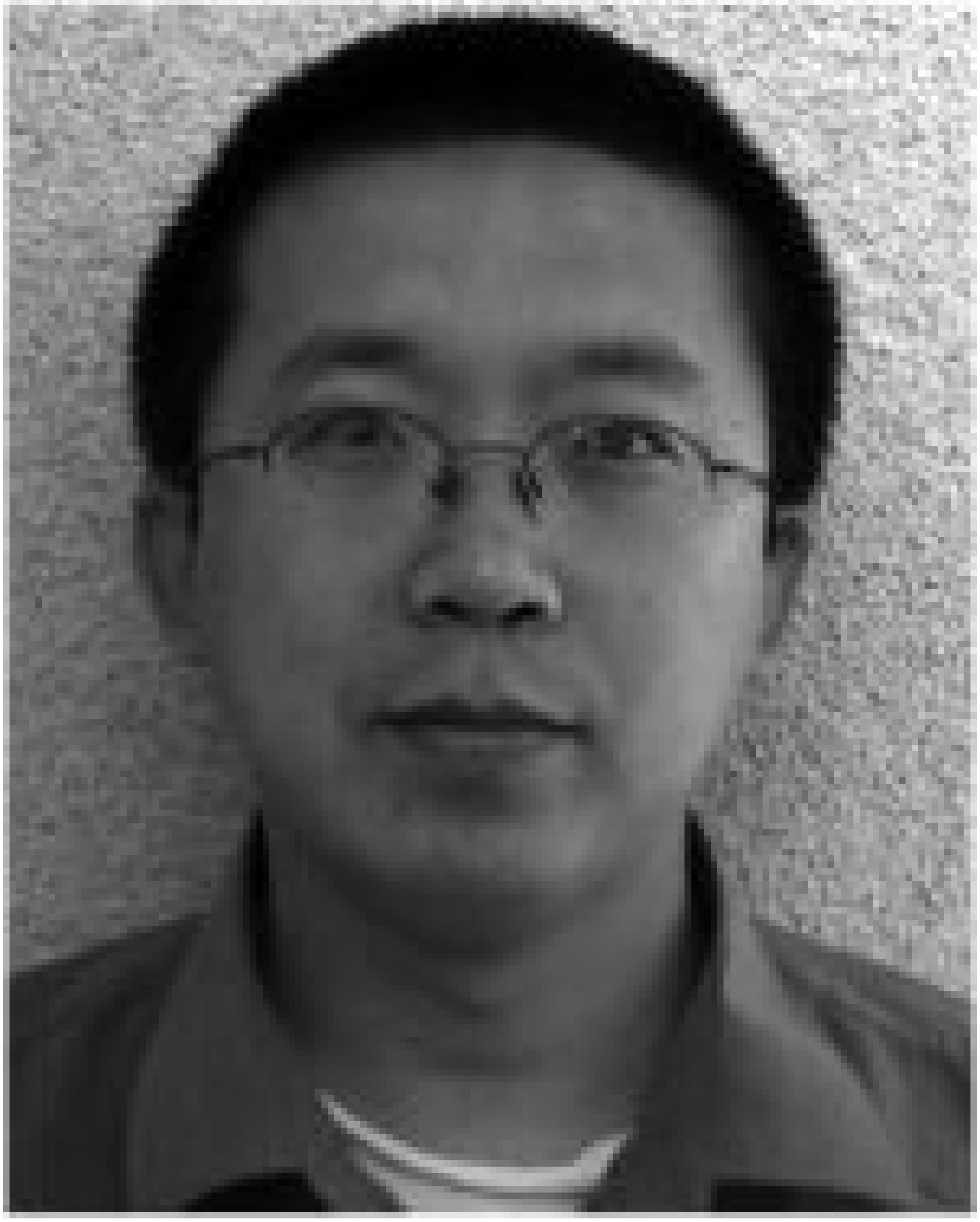
We described different types of noises which could render the syllable correlation envelope peak counting in the prone to error. Due to different characteristics of these noises, there is no one universal method to deal with all of them well. The approach we described in this paper was to design several different components, each addressing a specific subset of noise types. Finally, we tune the parameters and thresholds jointly to make these components work optimally through systematic multiparameter tuning. While removing a particular component from the system provided some insights into its relative effect on performance, such an approach does not ensure that the values of the other parameters are necessarily optimal. Further detailed experiments can help shed further light onto such details.

Evaluating the role of the estimates of speech rate derived in this work within specific application frameworks is outside the scope of the present work. Rate sensitive modeling in automatic speech recognition has been shown to provide performance improvements [54], and we expect that improved rate estimation to contribute toward improvement such models. Similarly, the results of the present work can contribute to other spoken language processing domains. In related work [45], acoustic measures of word prominence were shown to benefit from the algorithms presented in this paper. Further detailed application-specific evaluations of the proposed rate estimation remain as topics of future work.

Acknowledgments

This work was supported in part by grants from the Office of Naval Research (ONR), the National Science Foundation (NSF), and the U.S. Army.

Biography



Dagen Wang received the B.S. and M.S. degrees in electrical engineering from Peking University, Beijing, China, in 1997 and 2000, respectively, and the Ph.D degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2006.

He was with the Intel China Research Center, Beijing, China. He is now a Speech Scientist at the IBM T. J. Watson Research Center, Yorktown Heights, NY. His current research interest is on speech recognition and translation on limited-resource platforms. His general research interests include signal processing and artificial intelligence with applications to speech, language, and human–computer interaction problems.

Shrikanth S. Narayanan (S'88–M'95–SM'02) received the Ph.D. degree from the University of California, Los Angeles, in 1995.

He is Andrew J. Viterbi Professor of Engineering at the University of Southern California (USC), Los Angeles, where he holds appointments as Professor in electrical engineering and jointly in computer science, linguistics, and psychology. Prior to joining USC, he was with AT&T Bell Labs and AT&T Research, first as a Senior Member, and later as a Principal Member of its Technical Staff from 1995–2000. At USC, he is a member of the Signal and Image Processing Institute and a Research Area Director of the Integrated Media Systems Center, an NSF Engineering Research Center. He has published over 235 papers and has 14 granted/pending U.S. patents.

Dr. Narayanan is a recipient of an NSF CAREER Award, USC Engineering Junior Research Award, USC Electrical Engineering Northrop Grumman Research Award, a Provost Fellowship from the USC Center for Interdisciplinary Research, a Mellon Award for Excellence in Mentoring, and a recipient of a 2005 Best Paper Award from the IEEE Signal Processing Society. Papers by his students have won best student paper awards at ICSLP'02, ICASSP'05, and MMSP'06. He is an Editor for the *Computer Speech and Language Journal* (2007-present) and an Associate Editor for the IEEE *Signal Processing Magazine*. He was also an Associate Editor of the IEEE Transactions on Speech and Audio Processing (2000–2004). He serves on the Speech Processing and Multimedia Signal Processing technical committees of the IEEE Signal Processing Society and the Speech Communication Committee of the Acoustical Society of America. He is a Fellow of the Acoustical Society of America and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu.

References

[1]. Richardson, M.; Hwang, M.; AceroX, ADH. Improvements on speech recognition for fast talkers; Proc. Eurospeech; Budapest, Hungary. 1999; p. 411-414.

[2]. Greenberg, S. The switchboard transcription project; Tech. Rep., 1996 Johns Hopkins CLSP Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition; Baltimore, MD. 1997;

[3]. Beinum, FJK.; van Donzel, ME. Relationship between discourse structure and dynamic speech rate; Proc. Int. Conf. Spoken Lang. Process.; Philadelphia, PA. 1996; p. 1724-1727.

[4]. Mirghafori, N.; Fosler, E.; Morgan, N. Fast speakers in large vocabulary continuous speech recognition: Analysis and antidotes; Proc. Eurospeech'95; Madrid, Spain. 1995; Sep.. p. 491-494.

[5]. Martinez, F.; Tapias, D.; Alvarez, J. Towards speech rate independence in large vocabulary continuous speech recognition; Proc. ICASSP; Seattle, WA. 1998; May. p. 725-728.

[6]. Shriberg E, Stolcke A, Hakkani-Tur D, Tur G. Prosody-based automatic segmentation of speech into sentences and topics. Speech Commun. (Special Issue Accessing Information in Spoken Audio) 2000;32(1–2):127–154.

[7]. Byron, D.; Shriberg, E.; Stolcke, A. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues; Proc. Int. Conf. Spoken Lang. Process.; Denver, CO. 2002; p. 949-952.

[8]. Tamburini, F. Automatic prosodic prominence detection in speech using acoustic features: An unsupervised system; Proc. Eurospeech'03; Geneva, Switzerland. 2003; p. 129-132.

[9]. Morgan, N.; Fosler-Lussier, E. Combining multiple estimators of speaking rate; Proc. ICASSP; 1998; p. 729-732.

[10]. Nanjo, H.; Kawahara, T. Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition; Proc. ICASSP; 2002; p. 725-728.

[11]. Siegler, M. Measuring and compensating for the effects of speech rate in large vocabulary continuous speech recognition. Carnegie Melon Univ.; Pittsburgh, PA: 1995. Masters Rep.

[12]. Morgan, N.; Fosler, E.; Mirghafori, N. Speech recognition using on-line estimation of speaking rate; Proc. Eurospeech; Rhodes, Greece. 1997; p. 2079-2082.

[13]. Kitazawa, S.; Ichikawa, H.; Kobayashi, S.; Nishinuma, Y. Extraction and representation rhythmic components of spontaneous speech; Proc. Eurospeech; Rhodes, Greece. 1997; p. 641-644.

[14]. Speech filing system. [Online]. Available: http://www.phon.ucl.ac.uk/resource/sfs/

[15]. Pfau, T.; Faltlhauser, R.; Ruske, G. A combination of speaker normalization and speech rate normalization for automatic speech recognition; Proc. Int. Conf. Spoken Lang. Process.; Beijing, China. 2000; p. 362-365.

[16]. Nooteboom, S. The prosody of speech: Melody and rhythm. In: Hardcastle, IW.; Laver, J., editors. The Handbook of Phonetic Sciences. Blackwell; Oxford, U.K.: 1997. p. 640-673.

[17]. Poor, HV. An Introduction to Signal Detection and Estimation. Springer-Verlag; New York: 1985.

[18]. Narayanan, S.; Wang, D. Speech rate estimation via temporal correlation and selected sub-band correlation; Proc. ICASSP; Philadelphia, PA. 2005; Mar.. p. 413-416.

[19]. Stolcke A, Ries K, Coccaro N, Shriberg E, Bates R, Jurafsky D, Taylor P, Martin R, Van Ess-Dykema C, Meteer M. Dialogue act modeling for automatic tagging and recognition of conversational speech. Comput. Ling 2000;26(3):339–373.

[20]. Pfitzinger, HR.; Burger, S.; Heid, S. Syllable detection in read and spontaneous speech; Proc. ICSLP'96; Philadelphia, PA. 1996; p. 1261-1264.

[21]. Wang, D.; Narayanan, S. An unsupervised quantitative measure for word prominence in spontaneous speech; Proc. ICASSP; Philadelphia, PA. 2005; Mar.. p. 377-380.

[22]. Holmes JN. The JSRU channel vocoder. IEEE Proc. F. Commun, Radar Signal Process 1980;127 (1):53–60.

[23]. Jurafsky, D.; Bates, R.; Coccaro, N.; Martin, R.; Meteer, M.; Ries, K.; Shriberg, E.; Stolcke, A.; Taylor, P.; Ess-Dykema, CV. Automatic detection of discourse structure for speech recognition and understanding; Proc. IEEE Workshop Speech Recognition and Understanding; Santa Barbara, CA. 1997; Dec.. p. 88-95.

[24]. Zellner, B. Fast and slow speech rate: A characterisation for french; Proc. Int. Conf. Spoken Lang. Process.; Sydney, Australia. 1998; Dec.. p. 3159-3163.

[25]. Zheng, J.; Franco, H.; Stolcke, A. Rate-dependent acoustic modeling for large vocabulary conversational speech recognition; Proc. ISCA Tutorial and Research Workshop on Automatic Speech Recognition: Challenges for the New Millennium; Paris, France. 2000; p. 145-149.

[26]. O'Shaughnessy, D. Timing patterns in fluent and disfluent spontaneous speech; Proc. ICASSP; 1995; p. 600-603.

[27]. Weinstein C, McCandless S, Mondshein L, Zue V. A system for acoustic-phonetic analysis of continuous speech. IEEE Trans. Acoust. Speech, Signal Process Feb.;1975 ASSP-23(1):54–67.

[28]. Kasuya H, Wakita H. An approach to segmenting speech into vowel- and nonvowel-like syllables. IEEE Trans. Acoust. Speech, Signal Process Aug.;1979 ASSP-27(4):319–327.

[29]. Mermelstein P. Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Amer Oct.; 1975 58(4):880–883. [PubMed: 1194547]

[30]. Medress, M.; Diller, T.; Kloker, D.; Lutton, L.; Oredson, H.; Skinner, T. An automatic word spotting system for conversational speech; Proc. ICASSP'78; 1978; p. 712-717.

[31]. Rabiner L. On the application of energy contours to the recognition of connected word sequences. T Bell Labs Tech. J Nov.;1984 63(9):1981–1995.

[32]. Zwicker E, Terhardt EE, Paulus E. Automatic speech recognition using psychoacoustic models. J. Acoust. Soc. Amer Feb.;1979 65(2):487–498. [PubMed: 489818]

[33]. Reichl, W.; Ruske, G. Syllable segmentation of continuous speech with artificial neural networks; Proc. Eurospeech'93; Berlin, Germany. 1993; Sep.. p. 1771-1774.

[34]. Green, P.; Kew, N.; Miller, D. Speech representations in the SYLK recognition project. In: Cook, M.; Beet, S.; Crawford, M., editors. Visual Representations of Speech Signals. Wiley; New York: 1993.

[35]. Hermansky H, Morgan N. RASTA processing of speech. IEEE Trans. Speech Audio Process Oct.; 1994 2(4):578–589.

[36]. Fisher, W.; Zue, V.; Bernstein, V,J.; Pallet, D. J. Acoust. Soc. Amer. Suppl. A. Vol. 81. 1986. An acoustic-phonetic database; p. 592-600.

[37]. Roach, P. English Phonetics and Phonology. A Practical Course. Third ed.. Cambridge Univ. Press; Cambridge, U.K.: 2000.

[38]. Fisher, WM. The Spoken Natural Language Processing Group. National Inst. Standards Technol.; Gaithersburg, MD: 1997. Syllabification software.

[39]. Godfrey, J.; Holliman, E.; McDaniel, J. SWITCHBOARD: Telephone speech corpus for research and development; Proc. ICASSP'92; 1992; p. 517-520.

[40]. Handbook of the International Phonetic Association. Cambridge Univ. Press; Cambridge, U.K.: 1999.

[41]. Howitt, AW. Ph.D. dissertation. Mass. Inst. Technol.; Cambridge, MA: 2000. Automatic Syllable Detection for Vowel Landmarks.

[42]. Berthold, A.; Jameson, A.; Kay, J., editors. Interpreting symptoms of cognitive load in speech input; Proc. 7th Int. Conf. UM99, User Modeling; 1999; p. 235-244.

[43]. Garofolo, JS.; Lamel, LF.; Fisher, WM.; Fiscus, JG.; Pallett, DS.; Dahlgren, NL. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. Springer; Vienna, Austria: 1993.

[44]. Miller JL, Grosjean F, Concetta L. Articulation Rate and Its Variability in Spontaneous Speech: A Reanalysis and Some Implications. Phonetica 1984;41:215–225. [PubMed: 6535162]

[45]. Wang D, Narayanan S. An acoustic measure for word prominence in spontaneous speech. IEEE Trans. Speech, Audio, Language Process Feb.;2007 15(2):690–701.

[46]. Talkin, D. A robust algorithm for pitch tracking (RAPT); Proc. ICASSP; 1983; p. 1352-1355.

[47]. Young, I.; Gerbrands, J.; Vliet, L. v. Fundamentals of image processing. Delft Univ. Technol.; The Netherlands: 1998. [Online]. Available: http://www.ph.tn.tudelft.nl/Courses/FIP/noframes/fip.html

[48]. Bell A, Jurafsky D, Fosler-Lussier E, Gregory CGM, Gildea D. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. J. Acoust. Soc. Amer 2003;113:1001–1024. [PubMed: 12597194]

[49]. Fougeron C, Keating P. Articulatory strengthening at the edges of prosodic domains. J. Acoust. Soc. Amer 1997;101:3728–3740. [PubMed: 9193060]

[50]. Dauer R. Stress-timing and syllable-timing reanalyzed. J. Phonetics 1983;11:51–62.

[51]. Dauer, R. Phonetic and phonological components of language rhythm; Proc. Int. Congr. Phonetic Sci.; 1987; p. 447-450.

[52]. Clements, G. The sonority cycle and syllable organization. In: Dressler, W.; Luschutzky, H.; Pfeiffer, O.; Rennison, J., editors. Phonologica 1988. Cambridge Univ. Press; Cambridge, U.K.: 1992. p. 63-76.

[53]. Blevins, J. The syllable in phonological theory. In: Goldsmith, J., editor. Handbook of Phonological Theory. Blackwell; Oxford, U.K.: 1996. p. 35-59.

[54]. Mirghafori, N.; Fosler, E.; Morgan, N. Towards robust-ness to fast speech in ASR; Proc. ICASSP; 1996; p. 335-338.

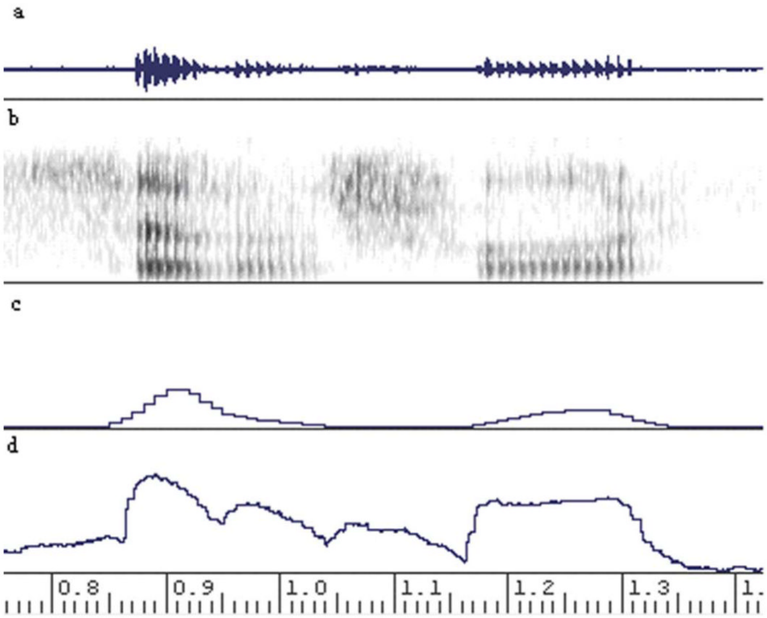


Fig. 1.
Sample speech utterance “SOME FORM” from the Switchboard corpus: (a) Speech waveform. (b) Wideband spectrum. (c) Correlation envelope (approach in this paper). (d) Wideband energy envelope.

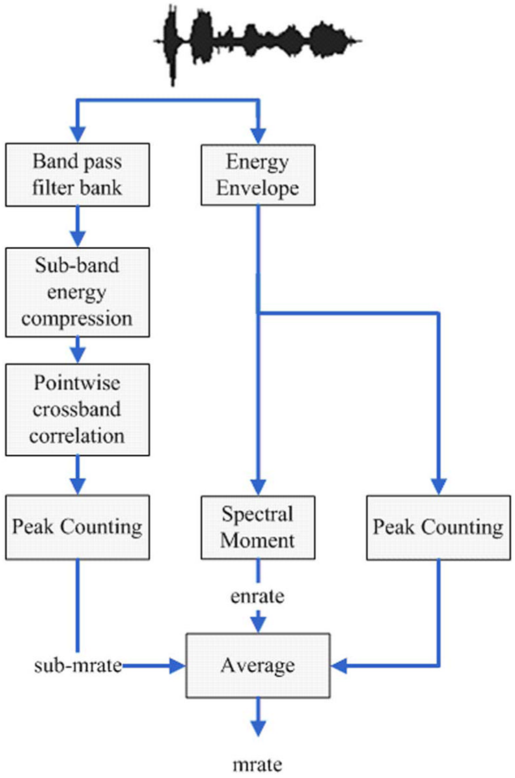


Fig. 2.

Major steps in computing “mrate” (adapted from [9]).

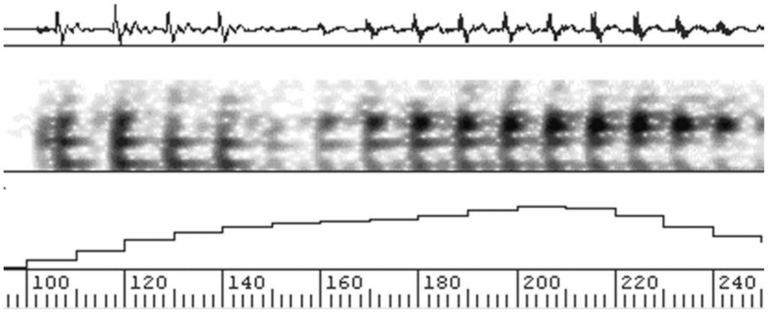


Fig. 3.

Illustration of peak smearing shown for the word “in-tro” (from the Switchboard corpus).

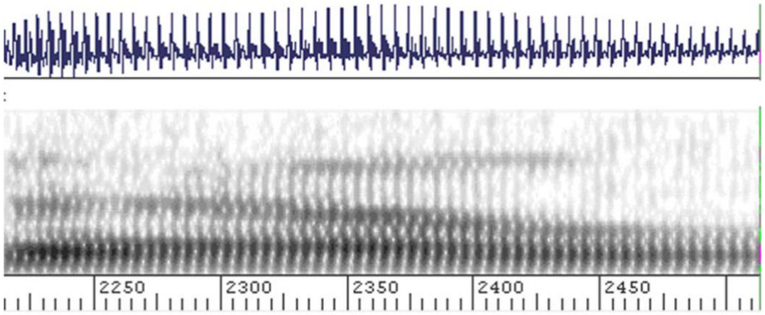


Fig. 4.
Overestimation for “So” (from Switchboard).

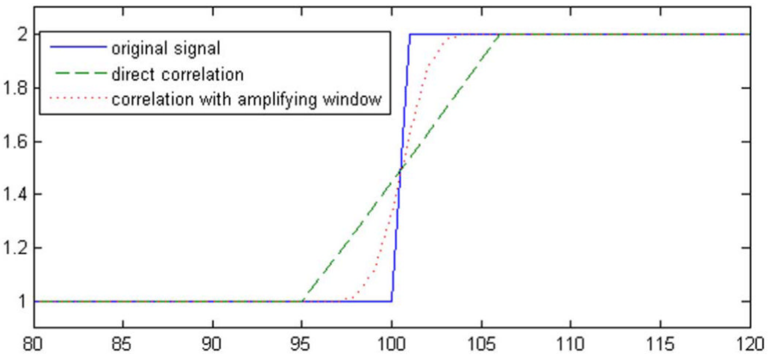


Fig. 5.
Weighting window effects for step functions. Correlation window length is set to 11, and the variance of Gaussian is 1.2.

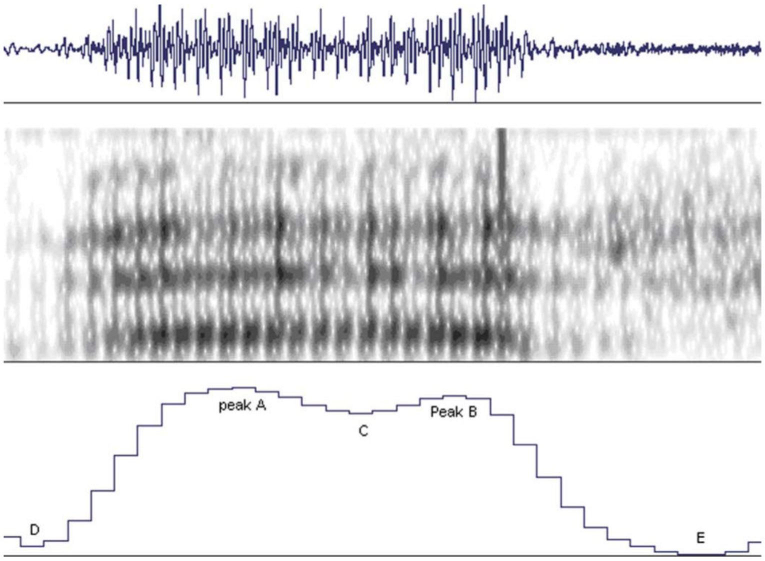


Fig. 6.
Syllable “BAD” in Switchboard 3994B.

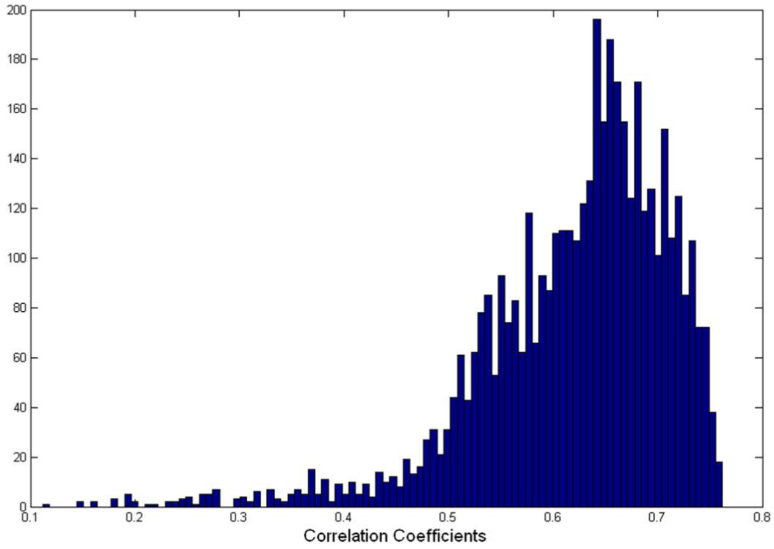


Fig. 7.

Monte Carlo simulation histogram.

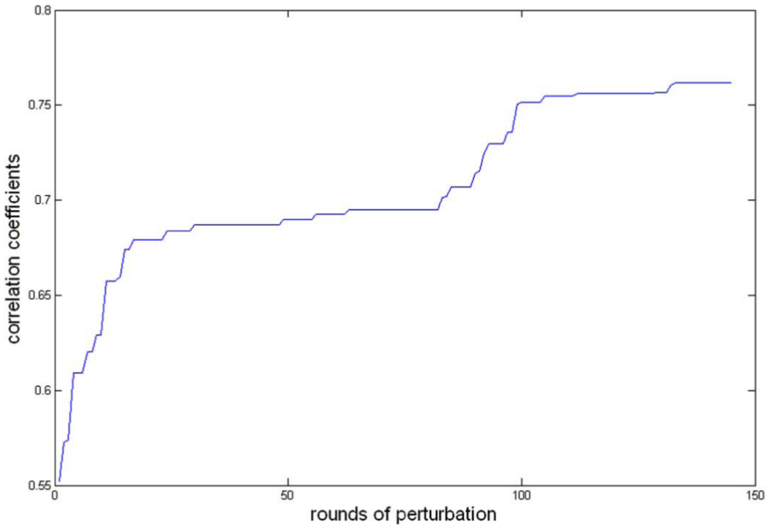


Fig. 8.

Perturbation yields monotonic improvement on correlation coefficient between test and reference data.

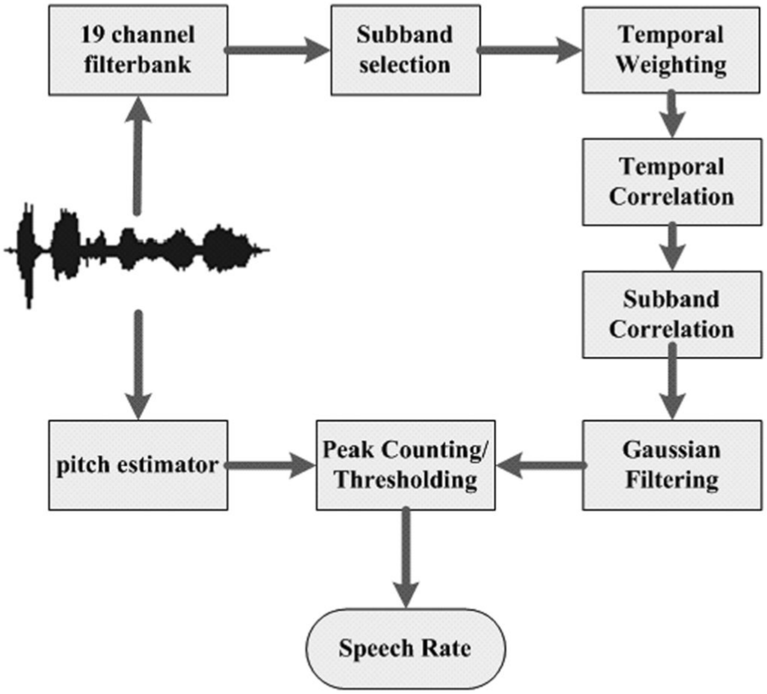


Fig. 9.

System flowchart for speech rate estimation.

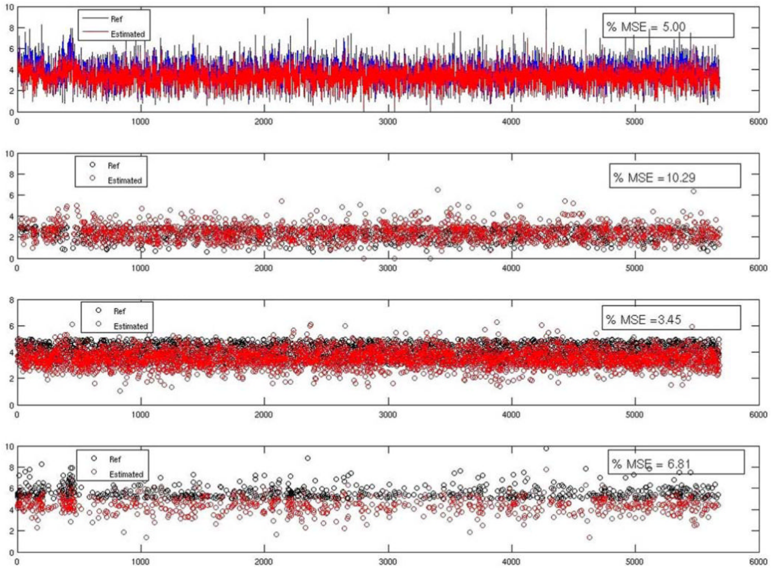


Fig. 10.

Estimated and reference rates for various data conditions (reference, blue; estimated, red). The top panel correspond to the results for the entire data, while the second, third, and last panels in the figure correspond to slow, normal, and fast speech, respectively. The horizontal axis is the ID of the spurts; the vertical axis is the MSE.

TABLE I

Optimal Parameter Settings

Value	Parameter
Temporal correlation window length (K)	11
Weighting Gaussian window variance	1.2
Number of selected sub-bands (M)	12
Smoothing window length	15
Smoothing Gaussian window variance	1.3
Neighboring peak distance threshold	13
Left-compare-only threshold	29

TABLE II

Experiments Run on the Development Set With Specific Components Removed From the Full System

MSE	Correlation	Mean error	Stddev error	Difference%	MSE%
Original: all components	0.690	0.242	0.849	0.0	5.32
All sub-bands included	0.687	0.257	0.851	0.43	5.39
No temporal correlation	0.654	−0.023	0.901	5.22	5.54
No Gaussian filtering	0.658	−0.018	0.896	4.64	5.48
No pitch validation	0.606	−0.280	0.962	12.17	6.85
No thresholding schemes	0.613	−0.190	0.953	11.16	6.44

TABLE III

Experimental Results Note: Enrate, Sub-Mrate, and Mrate are the Results From [9]

Measure	**Correlation**	**mean error**	**stddev error**
enrate	*.415*	*.747*	*1.405*
sub-mrate	*.637*	*.530*	*1.219*
mrate	*.671*	*.464*	*1.121*
Proposed Approach	**.745**	**.339**	**0.796**

TABLE IV

Effects of Syllable Nucleus Type on Speech Rate Estimation

Spurts type	# spurts	Correlation	Mean error
With vowels only	4346	0.737	0.322
With sonorants	1336	0.774	0.395
Combined data set	5682	0.745	0.339

TABLE V

Comparison to the Transcribed Syllable Location

Type	Correct %	Insertion %	Deletion %
Percent of total syllable number	80.6	3.8	15.6