SPEECH RATE ESTIMATION VIA TEMPORAL CORRELATION AND SELECTED SUB-BAND CORRELATION

Dagen Wang, Shrikanth Narayanan

Speech Analysis and Interpretation Lab
USC Viterbi School of Engineering

ABSTRACT

In this paper, we propose a novel method for speech rate estimation without requiring automatic speech recognition. It extends the methods of spectral subband correlation by including temporal correlation and the use of selecting prominent spectral subbands for correlation. Further more, to address some of the practical issues in previously published methods, we introduce some novel components into the algorithm such as the use of pitch confidence, magnifying window, relative peak measure and relative threshold. By selecting the parameters and thresholds from realistic development sets, this method achieves a 0.972 correlation coefficient on syllable number estimation and a 0.706 correlation on speech rate estimation. This result is about 6.9% improvement than current best single estimator and 3.5% improvement than current multi-estimator evaluated on the same switchboard database.

1. INTRODUCTION

Speech is a crucial component in human computer interaction. While tremendous progress has been made in automatic speech recognition, speech transcription -- which is the output of automatic speech recognition -- is far from providing all the information that one could retrieve from speech. For example, intonation, stress, timing, rhythm, and rate of speech all carry important information in speech and are crucial in speech perception. Inclusion of such information can facilitate better machine recognition and understanding of speech. Speech rate is one such key attribute. In this paper, we propose an algorithm for speech rate estimation.

1.1. Why rate of speech?

Speech rate has been initially investigated in the context acoustic modeling of speech recognition. It is apparent that the accuracy of a speech recognition system is severely affected when there are mismatches between the training and testing conditions. There are many possible factors causing these mismatches and speech rate is one of them [1]. Specifically, for better adapting to fast or slow speech, there has to be an estimation of speech rate. Only with this estimation could one select appropriate pre-trained acoustic models or adaptively set transition probabilities of the HMMs [4][5].

In recent years, with increasing interest in spontaneous speech recognition and interpretation, the role of speech rate estimates has become even more important. Research has found that local speech rate correlates with discourse structure. For example, global analysis of the discourse structure in paragraphs and clauses reveals that for each of the speakers 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 [3].

Local speech rate also 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 say slower local speaking rate, at those events [6][7]. Speech rate also correlates with prosodic prominence. Rate of speech detection and normalization has been found to be necessary in solving such problems [8].

1.2. How to measure speech rate?

It is quite natural for humans to use the term "fast", "normal", "slow" to describe speech rate. This classification has been applied in applications such as acoustic model selection [9] and HMM normalization [15]. However, this sort of classification is in itself fuzzy and needs humans to transcribe or manipulate. Practically, this classification can not be directly conveyed in the acoustic signal. So researchers in this area have adopted an intermediate quantitative measure of speech.

In most of the cases, speech rate is measured by counting phonetic elements per second. Words, syllables [9], stressed syllables, phonemes [10] are all possible candidates. However, it has been observed that humans do not follow strictly or consistently use these phonetic elements while control their speaking rate [11]. In some studies, the phone duration percentile, a comparison of measured versus expected phone duration, is shown to be robust with respect to lexical content and consistent with previous findings about the statistics of both long-term and short-term speech rate [11]. But for this method, the expected phone duration model can be well modeled only in very limited cases. For example, it cannot model large number of speakers or male and female speakers simultaneously.

Evidence from reiterative speech study [16] supports syllable to be a good estimate of speech rhythm, which is a similar measure to speech rate. Syllable is defined as a combination of elementary sounds uttered together with a single effort or impulse of the voice. Intuitively, syllables, by this definition, should have quite an even distribution under normal speed speech and their rate could be changed as a result of speech rate change. So it is used widely among speech rate researchers [6][9] [11]. In this work, we use syllable number per second as a measure of speech rate.

1.3. Previous work in speech rate estimation

1.3.1. With or without ASR?

Using automatic speech recognition to retrieve duration information about phonetic elements is straight forward. It is easy to get a phonetic alignment during speech recognition decoding and use this alignment timing information as a measure of speech rate [10]. It works well when speech recognition is reliable. But in the context of spontaneous speech, speech recognition is far from mature to robustly and precisely estimate these parameters. To address this issue at least partially, supervised alignment has been proposed. In cases where transcription is available, forced alignment can be used to provide better speech unit estimate. This method gives much more accuracy and has been successfully used in research [6][7]. However, such is not the case in the problem we are targeting in this work.

Moreover, one intended use of speech rate is to facilitate robust ASR in terms of appropriate model normalization and adaptation techniques. It is implied that speech rate estimation serves like a front end for speech recognition for a number of applications. Using speech recognition itself to address this problem is hence logically unsuitable. So it is quite natural to use the acoustic signal directly to study speech rate.

1.3.2. Acoustic study of speech rate

One classical way to get syllable count is through a full band spectrum/energy analysis and measures the dominant peak of the long-term envelope [13]. This however results in a lot of noise in the final curve and hence it is difficult to get syllable count robustly. This fact is apparent in Fig 1(d). The sample speech "some form" (from Switchboard) should only have 2 syllables, but (d) shows at least 4 dominant peaks. The results are hence not satisfactory.

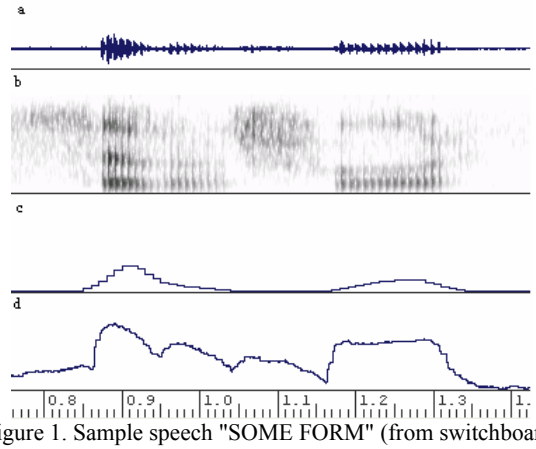

Figure 1. Sample speech "SOME FORM" (from switchboard) a) speech waveform b) wideband spectrum c) correlation envelope(approach in this paper) d) wideband energy envelope

As an alternate approach to the same problem, the first spectral moment of the broad-band energy envelope has been used as a speech rate measure [12]. While this method provided improved performance with conversational speech, it was however shown, using a one hour subset of the manually transcribed Switchboard data, the correlation between transcribed syllable rate and experiment rate was only about 0.4 (when both were measured over between-pause spurts) [12].

These two approaches assume that the wide-band energy peak as a valid representation for speech rate measure. A critical question then is how much information is lost or distorted in this process of using the wide-band energy curve, a lower dimension abstraction of the speech waveform. Specifically, are these losses and distortion crucial? From these aforementioned results, and supported by the example in Fig 1(d), the answer seems that this loss is indeed critical. For instance, the formant structures are lost in the wide band energy representation and this feature is crucial in fast speech syllable identification.

In [9], Morgan & Fosler-Lussier developed a sub-band based module that computes a trajectory that is the average product over all pairs of compressed sub-band energy trajectories. That is, if $x_i(n)$ is the compressed energy envelope of the i^{th} spectral band, a new trajectory y(n) is defined as:

$$y(n) = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} x_i(n)x_j(n) \qquad \text{(eq. 1)}$$

Where N is the number of bands, M=N(N-1)/2 is the number of unique pairs. By this method alone, correlation coefficients above 0.6 were achieved. Furthermore, it was shown in [9] that the performance would boost to 0.673 if multiple estimators were combined (wideband energy peak count, spectral moment count). It is apparent that this method addresses the formant structures we discussed earlier. By introducing a band wise correlation in the spectral domain, the syllable peak in the correlation curve gets boosted. But on the other hand, this algorithm does not address problems related to smoothness in the temporal domain.

The following sections discuss our approach. Our algorithm will generate a correlation envelope as shown in Fig 1(c).

2. FURTHER SPECIFIC ISSUES AND SOLUTIONS

In addition to the above-mentioned problem, there are several further issues that need to be tackled in designing a good speech rate estimator. Many of these are not well addressed in previous work. In the work proposed in this paper, we will further study the acoustic nature of speech and propose a set of algorithms to address different acoustic observations and related issues.

2.1. Background and consonant noise

In the region 0.78s-0.85s and 1.05s-1.15s of Figure 1, there are some apparent background noises. Such noises tend to introduce extra peaks in the final curve. Consonants, especially fricatives, also sometimes contribute extra peaks. We apply 2 methods to deal with this problem.

The first method is to use pitch (F0) information. When a peak is detected in a region with no voiced activity, it is rejected as noise. Since we do not care about the actual pitch value, it is helpful to use multi pitch estimators and fuse them together.

The other method is to use relative threshold to filter out the noise. "Relative" here means a scale with respect to the maximum peak. Like all threshold problems, it is dangerous to set the threshold value in a greedy fashion. Since we have the other approach to deal with the noise, we set the threshold rather low.

2.2. Energy curve smoothness

In all these methods, an energy curve is utilized. Like all short-time windowing methods, a larger window makes the curve

smoother yet loses fine details. A smaller window provides more detail but makes the curve noisy and in turn renders peak counting difficult.

In this paper, we propose a new method based on traditional windowing. Inspired by spectral cross correlation, and also by the fact that each syllable (i.e., similar spectral pattern) lasts for a while, we perform a cross correlation also in time domain. Let $\mathbf{x}_t, \mathbf{x}_{t+1} \ldots \mathbf{x}_{t+K-1}$ represent an ascending time order of sub-band energy vectors with length K. Then compute y_t as:

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

By this correlation, each syllable has a peak in its center, because it spans most of the part of this syllable. The parameter K is set by using a development test.

2.3. 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 a high number of syllables for that segment.

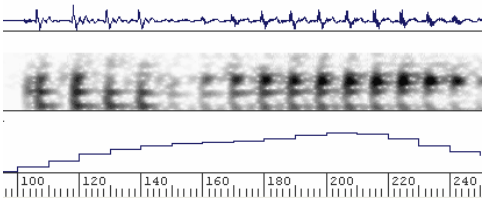

Figure 2. Illustration of peak smearing shown for the word "in-tro" (from switchboard corpus)

Figure 2 shows a smearing case where "in" and "tro" show only 1 peak. The reason is that the interval is smeared by the windowing and temporal correlation effect.

Let $w_0, w_1 \ldots w_{K-1}$ represent a serial of window coefficients. Perform a weighting operation on **x** first:

$$x_{t+j} = w_j x_{t+j} \quad \text{(eq. 3)}$$

Here we choose w as Gaussian window centered in the middle of the window. The reason for this choice is to amplify any discontinuities between neighboring syllables.

2.4 Over-estimation 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 method as proposed in [9].

As an example in figure3, "so" has only 1 syllable. For fixed sub-band, when one formant shifts from 1 band to another, it will generate one more peak.

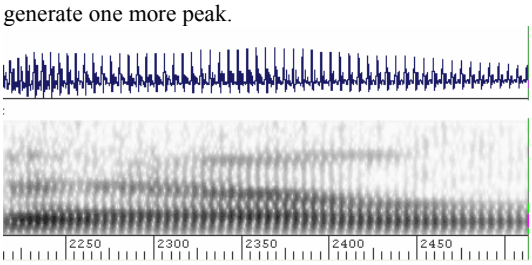

Figure3. Overestimation for "So" (from switchboard)

To address this issue, we propose a "selected sub-band correlation" method. First, instead of choosing only 4 sub-bands, we apply a 19 sub-band (as a facility provided in tool [14]). After getting $\mathbf{y_t}$, we choose the top M elements to do cross correlation as in [9]. By setting M optimally through the development test, the experiments show that it helps to resolve this issue successfully.

Another optimization relates to the relative peak measure. Each peak height is measured relative to the nearest largest minimum. For the extra peaks introduced by such formant movement, it always has a very low "height". By thresholding, such peaks could be removed.

3. ALGORITHM AND DESIGN

Inspired by these ideas, we implemented our full system according to the following steps:

First, the speech is passed through a 19-channel filter bank analyzer to get energy vector series. Second, the energy vectors are windowed and cross-correlated temporally. In the third step, result energy vector is cross-correlated in salient frequency bands. Finally, peak counting is performed on the final smoothed curve. Figure 4 gives a system flowchart.

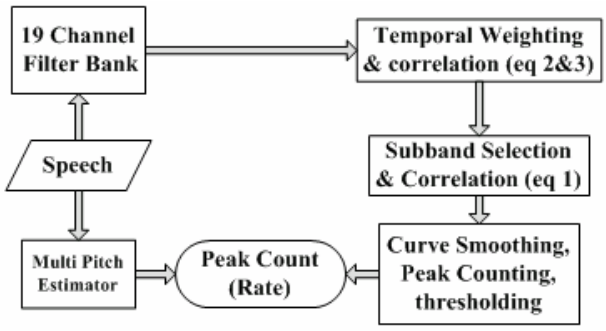

Figure4. System Flowchart

Here are some additional implementation comments:

1) The 19-channel filter bank analyzer uses two second-order section Butterworth band-pass filters [14]. Spaced as: 240 | 360 | 480 | 600 | 720 | 840 | 1000 | 1150 | 1300 | 1450 | 1600 | 1800 | 2000 | 2200 | 2400 | 2700 | 3000 | 3300 | 3750

2) We apply 2 pitch estimators: ESPS get_f0 call and cepstrum based estimation [14], use the union of the two as the pitch estimate.

3) For curve smoothing, we apply a Gaussian filter.

4. EXPERIMENTS AND RESULT

We use the same switchboard database and similar evaluation methods as in [9]. A total of 5565 spurts (all that we had in hand) were phonetically hand transcribed by linguists in the Switchboard Transcription Project at ICSI [2]. A transcribed syllable rate was computed by dividing the number of syllables occurring in the region by the length of the spurt. Similarly, we treat this rate as a reference rate. We use the detected rate to correlate with the reference rate to get the final agreement measure.

The ideas and algorithm described above have a heavy heuristic flavor based on analysis of spontaneous speech. At this stage the approach is not set up as a straight machine learning approach

where transcribed data help to setup optimal statistical models. We argue that, in fact, such learning ideas can benefit when we know what (and, how) "feature" correlates with the subjects' production. We believe that the "model" we setup here provides a step in the direction of helping handle the complexities underlying processing spontaneous speech.

The biggest challenge comes from the setting of many parameters that exhibit complex, and often confounding, correlations between one another. We address the issue through the following methodology:

Firstly, we are trying to group the parameters such that each group is independent or has little correlation with the others. The purpose of this step is to reduce the parameters' dimensionality such that a big complex problem can be divided into some small relatively simplified problems. In our experiment, the temporal correlating parameters, the spectral correlating parameters, the smoothing parameters/ peak counting thresholds are the 3 major groups. We normally fix the other 2 groups in an acceptable range and tune the current group's parameter using the development set.

Secondly, we do a sensitivity analysis wherein we pay close attention to the parameters that are quite sensitive relative to those that are not that influential to the final performance. For example, the temporal correlation window length ("K" in sec 2.2) was found to be sensitive and needed detailed experiments to set up. On the contrary, the temporal weighting parameters were less sensitive and relatively easy to setup.

Thirdly, by carefully inspecting the data we can set reasonable bounds on parameter selection. For example, the count of selected subband ("M" in sec 2.4) should have a close relation with formants numbers. So we only consider the range of 3 and slightly larger. This is a great reduction from the original 19 bands.

Lastly, we randomly select 315 spurts as a development set and used this to directly set the parameters of the algorithm. Of course, this is based on the good design of the previous steps. We always run several rounds of cross-validation until we reach a local maximum.

With all these efforts, we achieve the following result in Table 1. Besides speech rate, we also measure the correlation coefficients between the reference and detected syllable numbers.

Measure	Correlation	Mean error	Std error
Syllable #	0.972	1.143	2.257
Rate of speech	0.706	0.340	0.848

Table 1. Results table

This result is about 6.9% improvement than single estimator and 3.5% improvement than multi-estimator evaluated on the same database in [9].

For the envelope, an example is provided in fig 1(c).

5. CONCLUSIONS

Experiments have shown that the method proposed in this paper offer further advantages over previous methods. For instance, the envelope output is smoother and has better syllable count performance than previous methods. (fig 1(c)).

Based on the description in Sec 4, it is clear that the parameter selections are empirical and not guaranteed to by formal optimization. Even though we get fairly good performance, we still believe there is a great potential to further boost the performance. A possible alternative approach would be designing an adaptive algorithm for dynamic parameter adjustment.

6. REFERENCES

[1] Matthew Richardson, Mei-Yuh Hwang, Alex Acero, Xuedong Huang, "Improvements On Speech Recogniton For Fast Talkers", *Eurospeech*, 1999

[2] Greenberg, S. "The Switchboard Transcription Project", in F. Jelinek, editor, *1996 LargeVocabulary Continuous Speech Recognition Summer ResearchWorkshop* Technical Reports

[3] Koopmans-van Beinum, et al, "Relationship between discourse structure and dynamic speech rate". In: H.T. Bunnell & W. Idsardi (Eds), *Proceedings ICSLP96*.

[4] Mirghafori N., Fosler E., and Morgan N. "Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes". *Proceedings of the Eurospeech Conference*, Madrid, pp. 491-494. Sep, 1995

[5] Martinez F., Tapias D. and Alvarez J. "Towards Speech Rate Independence in Large Vocabulary Continuous Speech Recognition". *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle. pp.725-728. May 1998.

[6] E. Shriberg, A. Stolcke, D. Hakkani-Tur, G. Tur, "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics", *Speech Communication* 32(1-2), 127-154 (Special Issue on Accessing Information in Spoken Audio), 2000

[7] D. Baron, E. Shriberg, and A. Stolcke, "Automatic Punctuation and Disfluency Detection in Multi-Party Meetings Using Prosodic and Lexical Cues". *Proc. Intl. Conf. on Spoken Language Processing*, Denver, vol. 2, pp. 949-952, 2002

[8] Tamburini F, "Automatic Prosodic Prominence Detection in Speech using Acoustic Features: an Unsupervised System". In Proc. *Eurospeech* 2003, Geneva, 129-132

[9] N. Morgan and E. Fosler-Lussier. "Combining multiple estimators of speaking rate". In IEEE *ICASSP-98*, Seattle, WA, May 1998

[10] H.Nanjo and T.Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition", In *Proc. IEEE-ICASSP*, pp.725--728, 2002.

[11] Matthew A. Siegler, "Measuring and Compensating for the Effects of Speech Rate in Large Vocabulary Continuous Speech Recognition", Masters Report, Carnegie Mellon University, 1995

[12] Morgan, N., Fosler, E., and Mirghafori, N., "Speech Recognition using On-line Estimation of Speaking Rate", *EUROSPEECH* 1997, pp 2079-2082,Greece, 1997.

[13] Kitazawa, S., Ichikawa, H., Kobayashi, S., and Nishinuma, Y., "Extraction and Representation Rhythmic Components of Spontaneous Speech", *EUROSPEECH* 1997, pp. 641-644.

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

[15] Thilo Pfau, Robert Faltlhauser, Gunther Ruske, "A Combination of Speaker Normalization and Speech Rate Normalization for Automatic Speech Recognition", *ICSLP* 2000

[16] Nooteboom, S. (1997). "The prosody of speech: melody and rhythm", in W.J. Hardcastle, J. Laver (eds.) *The handbook of phonetic sciences*, Blackwell, Oxford, 640-673.