

A comparative study of speech rate estimation techniques

Tomas Dekens¹, Mike Demol¹, Werner Verhelst¹, Piet Verhoeve²

¹Vrije Universiteit Brussel, dept. ETRO-DSSP, Pleinlaan 2, B-1050 Brussels, Belgium

²Corporate R&D dept., TELEVIC nv, Leo Bekaertlaan 1, B-8870 Izegem, Belgium

{tdekens, midemol, wverhels}@etro.vub.ac.be, p.verhoeve@televic.com

Abstract

In this paper we evaluate the performance of 8 different speech rate estimators [1, 2, 3, 4, 5] previously described in the literature by applying them on a multilingual test database [6]. All the estimators show an underestimation at high speech rates and some also suffer from an overestimation at low speech rates. Overall the tested methods obtain high correlation coefficients with the reference speech rate. The Temporal Correlation and Selected Sub-band Correlation method (tcssbc), which uses sub-band and time domain correlation for detecting the number of vowels or diphthongs present in the speech signal, shows little errors and appears to be the most appropriate overall technique for speech rate estimation.

Index terms: cross lingual comparison, speech rate estimation

1 Introduction

Knowledge of the speech rate is crucial for a lot of different applications. For instance, in automatic speech recognition (ASR) differences in speech rate may cause mismatches between training and testing conditions and can significantly deteriorate the recognition accuracy [7]. If an estimation of the speech rate is known beforehand, it is possible to select a suitable pre-trained acoustic model or to adapt the transition probabilities of the used Hidden Markov Models [8, 9].

A lot of research on speech rate estimation has already been done. Unfortunately, all the developed speech rate estimation methods have been tested on different database material which makes it difficult to compare the different estimators in the different studies. Also, to our knowledge, no comparison has been made that test the performance of these methods on the same database. So, in this paper we give an overview of several speech rate estimation methods found in literature [1, 2, 3, 4, 5] and investigate their performance on the same multilingual database [6]. In total we compare 8 different estimators. All methods use acoustic measurements of the speech signal to derive the rate of speech (ROS). Also, all except one (*enrate*) rely on detecting the number of vowels or diphthongs present in the speech signal. As every syllable contains only one vowel or diphthong, the

number of syllables per second is used as a measure for the speech rate.

In the following section we will briefly describe the different speech rate estimators. In section 3, the test data used to evaluate the estimators is described, followed by a discussion of the results obtained with the different methods. Finally, we draw our conclusions in section 4.

2 Speech rate estimators

The first method that we will discuss is *enrate* [1]. *Enrate* is the first spectral moment of the broadband energy envelope, and is defined as follows:

$$enrate = \frac{\sum_{k=s}^L k |X(k)|^2}{\sum_{k=s}^L |X(k)|^2} \quad (1)$$

where $X(k)$ represents the Fourier transform of the windowed energy envelope of the speech signal. To obtain the energy envelope the signal is half-wave rectified, low-pass filtered and down sampled to 100Hz. Start point s in Eq. 1 is chosen in such a way that the effect of the DC content in the envelope is reduced (e.g. 0.5 Hz), and endpoint L is chosen to correspond to 16Hz. When the speech rate is high the energy envelope will possess more energy at high frequencies. As a result, more weight will be assigned to higher values of k and the *enrate* will increase, see Eq. 1. Thus, we can expect a certain correlation between the real speech rate and the *enrate*. Note that with *enrate* the result is not in syllables per second, but in Hz.

Based on the *enrate* an enhanced method, called *mrate*, was proposed in [2]. *Mrate* is the average of *enrate* and two different peak counting estimators. The first peak counting is done on the energy envelope of the signal and the result is also a direct estimation of the speech rate by itself. We will call this estimator *peakrate*. The second peak counting is executed on the point-wise correlation y (see Eq. 2) of the energy envelope of 4 different band pass signals (with band edges: 300, 800, 1500, 2500, 4000 Hz).

$$y(n) = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^N x_i(n) x_j(n) \quad (2)$$

where x_i represents the energy envelope of the i^{th} sub-band signal. $M = N(N-1)/2$ is the number of unique pairs and $N (=4)$ is the number of sub-bands. The peak

counting on y also represents the detected number of vowels or diphthongs in the speech signal and will be called *mpeakrate*. So *mrate* will be the average of *enrate*, *peakrate* and *mpeakrate*.

Additionally, we could also apply a peak counting solely on the energy envelope of the first sub-band signal, thus corresponding to the band (300, 800) Hz. We will call this estimator *peakfrate*.

The *tcssbc* method (Temporal Correlation and Selected Sub-band Correlation), described in [3], performs a time domain cross correlation prior to the point wise correlation between the energy envelopes of 18 sub-band signals (with band-pass edges: 240, 360, 480, 600, 720, 840, 1000, 1150, 1300, 1450, 1600, 1800, 2000, 2200, 2400, 2700, 3000, 3300, 3750 Hz). $y_i(n)$ is then defined as:

$$y_i(n) = \frac{1}{K(K-1)} \sum_{j=0}^{K-2} \sum_{p=j+1}^{K-1} x_i^{wm}(n+j)x_i^{wm}(n+p) \quad (3)$$

where x_i^{wm} represents the envelope of the i^{th} sub-band signal after windowing with a Gaussian window of length K and with its origin at position n . The time domain correlation will smooth the energy envelopes, on which then the sub-band correlation (see Eq. 2) can be applied. Counting the number of peaks will again lead to an estimation of the speech rate. The value of the parameters K and N in equations 2 and 3 for *tcssbc* are set during an optimization step to 8 and 11, respectively.

Another method that we evaluated, calculates the *vowel strength* (VS), and tries to detect the vowel onset points (VOPs) [4]. By using subharmonic summation [10], the instantaneous pitch is determined and a combined strength of spectral measurements is calculated in one pitch period. Weighting this combined strength with the maximum of the subharmonic sum spectrum results in the VS. The VS is then differentiated and the peaks of the resulting signal are candidates for the VOPs. Some heuristic rules are applied to detect the actual VOPs. We will call this speech rate estimator *vorate*.

The last estimator uses linear prediction (LP) to detect the VOPs [5]. First, the Hilbert envelope of the LP residue of the speech signal is computed. Next, the Hilbert envelope is convolved with a Gabor window in order to obtain the *VOP evidence plot*. The peaks of the *VOP evidence plot* are considered to be potential VOPs. Finally, some heuristic rules are applied to determine which peaks correspond to a VOP. We call this estimator *voprate*.

3 Experiments and results

3.1 Testdata

We applied the different estimators on a multilingual database [6] to evaluate their performance. The database consists of 552 sound files covering 6 different European languages. In every language speakers were asked to utter 8 different utterances at slow, normal and

fast speech rates. The different languages and the number of speakers are Dutch (4), English (4), French (5), Italian (3), Romanian (4) and Spanish (3). The utterances have an average duration of 28 seconds and the average number of syllables per utterance is 115. The speech is sampled at 44100 Hz. As the number of syllables in the uttered sentences is known, we were able to calculate the reference speech rate in syllables per second for every utterance.

The Dutch test data were split up into two parts; 60 sound files were used to optimize the different estimators, the remaining 36 files were used as test files. During a manual optimization step, the parameters of the different methods were set such that a compromise was reached between a high correlation with the reference speech rates and a low mean square error (MSE). This tradeoff was determined in an empirical fashion.

3.2 Results

We will discuss and compare the results from the Dutch and the French data sets. This gives us the opportunity to investigate the performance of the estimators in different languages and at higher speech rates, as the French speakers tended to speak faster than the Dutch ones. This difference in speed was clearly visible on the histograms of the speech rates of both languages. We used the same optimized parameters from the Dutch subset for the other languages as well.

Table 1 and Table 2 show the results for Dutch and French, respectively. Figure 1 and Figure 2 illustrate the estimated versus the reference speech rate. The line in the plots indicates where the estimated and reference speech rates are equal. *Enrate* doesn't produce a result directly in syllables per second and therefore it is difficult to compare its result with the reference speech rate. *Enrate* has also the lowest correlation coefficient of the estimators considered. These 2 major drawbacks make *enrate* less appropriate for speech rate estimation.

An overestimation at low speech rates is observed especially with *vorate* and *voprate* (see Fig 1). At slow speech rates peaks from other phones become larger and could be mistakenly counted as vowel. *Mrate* and *voprate* also show an underestimation at high rates. This underestimation becomes more pronounced when we consider the results for the French language. At high speech rates all estimators show an underestimation of the actual speech rate. All the methods (except *enrate*) rely on the counting of maxima at places where vowels are expected to occur. At high speech rates, more co-articulation exists and the peaks tend to merge together causing an underestimation. In most methods, the merging of the peaks is also re-enforced by the presence of voiced consonants like the /m/ and /n/. For the *tcssbc* and *mpeakrate* methods, this underestimation appears less pronounced. Both methods use a sub-band approach in combination with a cross-band correlation. The voiced consonants like /m/ possess most of their energy in the lowest sub-band only, and therefore will not cause a peak in the cross correlation. As a result

their influence on the merging of peaks is reduced and better results are obtained.

Mpeakrate has less underestimation than the other methods, but on the other hand also has a lower correlation with the reference speech rate, see Table 1. The tcssbc method has a high correlation and shows small differences with the reference speech rate (small MSE) compared to the other estimators. The tcssbc method can be considered to perform best of all the speech rate estimators studied.

The performance and the results of the estimators were similar for the other languages as well. For languages with a speech rate distribution similar to the one observed for the French language (Italian, Romanian and Spanish) an underestimation at the higher speech rates was also noticed. For English the same conclusion could be drawn as for Dutch. So, the performance of the estimators is not directly language dependent. The language dependency only results from the fact that the performance of the estimators degrades at high speech rates.

A final remark could be made concerning the reference speech rate. As mentioned above this reference is derived from the knowledge of the number of syllables in the test data. In some cases however, the speaker added syllables (especially ‘schwa’) when speaking slowly (e.g., for Dutch: ‘firrema’ instead of ‘firma’). This effect of adding new verbal material is also described in [11]. When speaking at a high rate speakers tended to omit syllables or small words (e.g. ‘difficult to sleep’ sometimes became ‘difficult sleep’). These effects are difficult to detect reliably and were not taken into account for calculating the reference speech rate, instead, the number of syllables the speakers were supposed to utter was used. The adding of syllables occurs very rarely and its influence on the results is negligible. Omitting sounds on the other hand happens more frequently. Especially in French at reference rates of more than 6 syllables per second, an error in the reference speech rate of about 0.3 syllables per second should be considered. This effect also contributes to the observed underestimation, but not to the same extent as the peak merging effect.

	enrate	peakrate	mpeakrate	mrate	peakfrate	tcssbc	vorate	voprate
Correlation	0.8943	0.9451	0.8974	0.9502	0.9357	0.9331	0.9142	0.9141
Mean error	-0.8598	-0.0666	0.2389	-0.2844	-0.0838	-0.0186	0.0810	-0.1921
Stddev error	1.0726	0.3673	0.4030	0.4367	0.3649	0.3718	0.4452	0.4881

Table 1: Correlation coefficient, mean error and standard deviation error for the different estimators (Dutch)

	enrate	peakrate	mpeakrate	mrate	peakfrate	tcssbc	vorate	voprate
Correlation	0.8047	0.9029	0.8419	0.9323	0.9330	0.9242	0.9212	0.9321
Mean error	-1.4319	-0.5275	-0.1680	-0.7091	-0.6248	-0.1884	-0.3121	-0.7059
Stddev error	1.5896	0.5172	0.5589	0.5291	0.4645	0.4659	0.5067	0.5344

Table 2: Correlation coefficient, mean error and standard deviation error for the different estimators (French)

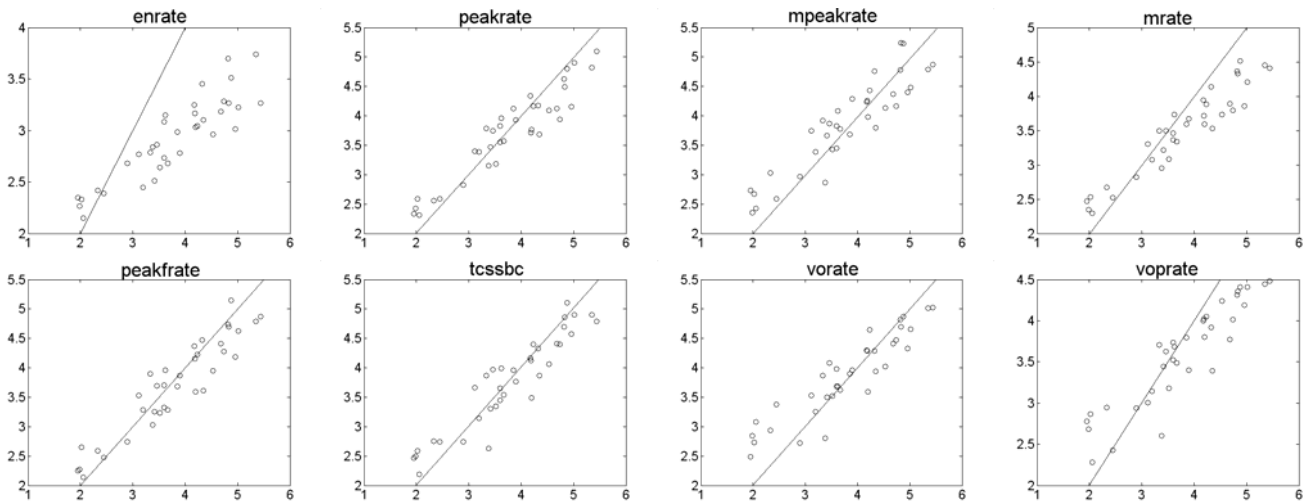


Figure 1: Results obtained with different estimators on Dutch speech. The line in the plots shows where estimated and reference speech rates are equal. X-axis: reference speech rate, Y-axis: estimated speech rate.

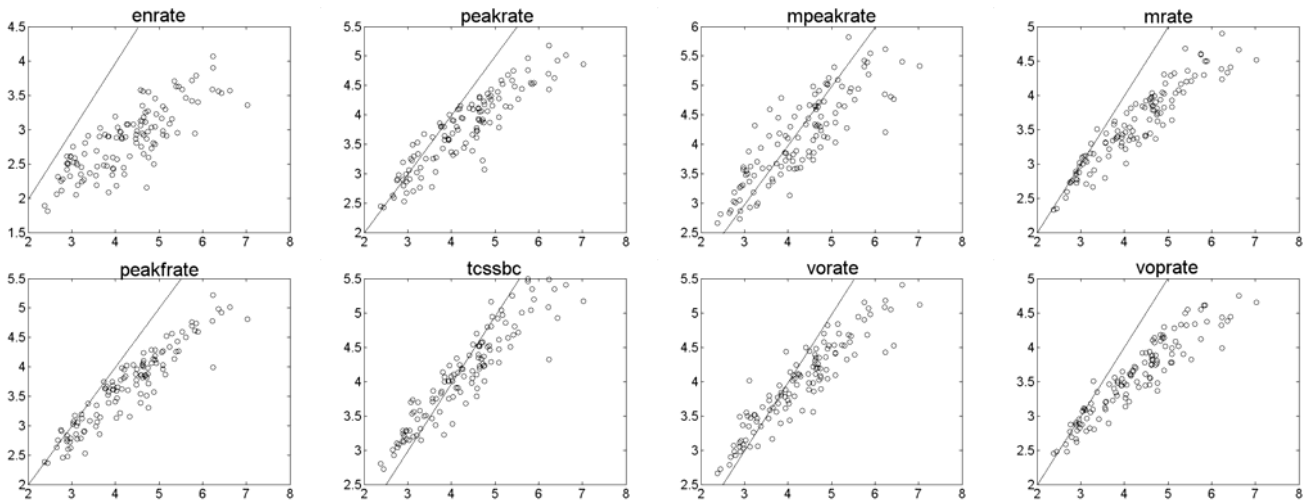


Figure 2: Results obtained with different estimators on French speech. The line in the plots shows where estimated and reference speech rates are equal. X-axis: reference speech rate, Y-axis: estimated speech rate.

4 Conclusions

In this paper we have evaluated 8 different methods for speech rate estimation. All estimators except for *enrate* use peak counting to detect the position of the vowels in the utterance. All methods attain high correlations with the reference speech rate. At slow speech rates *vorate* and *voprate* show an overestimation of the reference speech rate. At high speech rates almost every estimator shows a rather strong underestimation. This is mainly due to the merging of peaks caused by co-articulation and voiced consonants. The *tcssbc* and *mpeakrate* methods suffer less from peak merging and thus they show a clearly smaller underestimation. The *tcssbc* method seems, from our test, the most suitable for estimating the speech rate (high correlation and low MSE).

5 Acknowledgments

Parts of the research reported on in this paper were supported by the Institute for the Encouragement of Innovation through Science and Technology in Flanders (IWT) through the research grant of Mike Demol and IWT projects O&O/040803 and SMS4PA (IWT040803).

6 References

- [1] N. Morgan, E. Fosler, and N. Mirghafori, "Speech Recognition using On-line Estimation of Speaking Rate", Eurospeech, Rhodes, Greece, pp. 2079-2082, September 1997.
- [2] N. Morgan, and E. Fosler-Lussier, "Combining multiple estimators of speaking rate", IEEE ICASSP, Seattle, WA, pp. 729-732, May 1998.
- [3] D. Wang, and S. Narayanan, "Speech Rate Estimation via Temporal Correlation and Selected Subband Correlation", IEEE ICASSP, Philadelphia, PA, pp. 413-416, March 2005.
- [4] D.J. Hermes, "Vowel-onset detection", Journal of the Acoustical Society of America 87, pp. 866-873, 1990.
- [5] S.R.M. Prasanna, J.M. Zarachiah, and B. Yegnanarayana, Workshop on Spoken Language Processing, TIFR, Mumbai, January 9-11, 2003.
- [6] M. Demol, W. Verhelst, and P. Verhoeve, "A Study of Speech Pauses for Multilingual Time-Scaling Applications", in proc. ISCA-ITRW Multiling 2006, Stellenbosch, South Africa, April 9-11, 2006.
- [7] M. Richardson, M-Y. Hwang, A. Acero, and X. Huang, "Improvements On Speech Recognition For Fast Talkers", Eurospeech, Budapest, Hungary, September 5-9, 1999.
- [8] N. Mirghafori, E. Fosler, and N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes", Proceedings of the Eurospeech Conference, Madrid, pp. 491-494, September 1995.
- [9] F. Martinez, D. Tapias, and J. Alvarez, "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.
- [10] D.J. Hermes, "Measurement of pitch by subharmonic summation", Journal of the Acoustical Society of America 83, pp. 257-264, 1988.
- [11] B. Zellner, "Fast and Slow Speech Rate: a Characterisation for French", In *Proceedings ICSLP, 5th International Conference on Spoken Language Processing*, 7, Sydney, Australia, pp. 3159-3163, 1998.