ESTIMATING THE SPEAKING RATE BY VOWEL DETECTION

T.Pfau, G.Ruske

Institute for Human-Machine-Communication, Technical University of Munich, Arcisstr. 21, 80290 München
Tel.: +49 89 289-28554, Fax: +49 89 289-28535, e-mail: {pfa,rus}@mmk.e-technik.tu-muenchen.de

ABSTRACT

We present a new feature-based method for estimating the speaking rate by detecting vowels in continuous speech. The features used are the modified loudness and the zerocrossing rate which are both calculated in the standard preprocessing unit of our speech recognition system. As vowels in general correspond to syllable nuclei, the feature-based vowel rate is comparable to an estimate of the lexically-based syllable rate. The vowel detector presented is tested on the spontaneously spoken German Verbmobil task and is evaluated using manually transcribed data. The lowest vowel error rate (including insertions) on the defined test set is 22,72% on average over all vowels. Additionally correlation coefficients between our estimates and reference rates are calculated. These coefficients reach up to 0,796 and therefore are comparable to those for lexically-based measures (like the phone rate) on other tasks. The accuracy is sufficient to use our measurement for speaking rate adaptation.

1. INTRODUCTION

A main problem for speaker independent automatic speech recognition systems is the variability of the speech signal. The same sequence of words uttered by different speakers or even uttered several times by one speaker never results in identical speech signals.

Recent research has shown efficient methods for compensating the inter-speaker variability by either adaptation to the individual speaker's characteristics (e.g. MAP and MLLR related approaches) or by speaker normalization of the speech signal (e.g. vocal tract normalization).

Various reasons - among those the rate of speech (ROS) should be mentioned as very important - can be held responsible for the effect of intra-speaker variability of the speech signal. Increased coarticulation effects as well as the use of different pronunciation variants can be observed at higher speech rates. In first experiments on the German Verbmobil database we were able to confirm the results of [3] and [4], which showed a degradation in performance of automatic speech recognition systems for exceptionally fast or slowly spoken sentences. In figure 1 the word error rates on the Verbmobil crossvalidation set 1996 are shown (the set was divided into 12 bins according to the speech rate measured for each utterance and then the error rates for each bin were determined).

Lately some approaches have been made to compensate for the effects of different speaking rates. In [2], [3] and [4] changes of the HMM state transition probabilities improved the performance of the systems used as well as changes of a neural net (MLP) phonetic probability estimator in [2] and [3].

For being able to apply such compensation techniques either the speech rate must be determined a priori or at least an estimate of the speech rate must be available. Therefore our work described in this paper concentrates on a new approach for robust online estimation of the ROS.

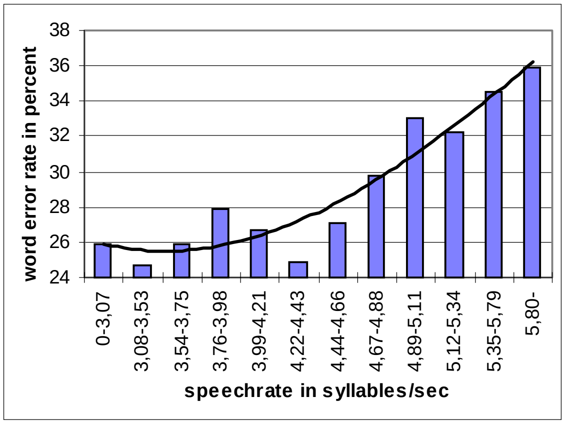

Figure 1: Word error rates on the Verbmobil crossvalidation set 1996

Good and robust measures of the speech rate can be found when counting phonetic units after the recognition process [2]. However, for an online compensation of the ROS a recogniton independent measure has to be found. The estimate of the ROS should be derived concurrently to the recognition process in order to be able to use the speech rate as an additional knowledge source to improve recognition results. For example in [5] a MLP is applied for estimating the ROS and in [1] a simple estimator of the speaking rate directly based on the speech signal is presented, which uses the energy envelope of the speech signal as the prominent feature.

Our new measure presented here is based on the detection of vowels in the speech signal by utilizing a special feature, the so-called „modified loudness“, and additionally the zero-crossing rate. By this method the estimation of the ROS can be performed simultaneously to the recognition process. Here we investigated the efficiency of this ROS detector alone. Furthermore it should be mentioned that our method can easily be combined with usual ROS measurements derived from phoneme recognition results.

2. USING THE MODIFIED LOUDNESS

Vowels in general correspond to the syllable nuclei of speech and counting vowels therefore should be strongly related to counting syllables. Furthermore the sequence of vowels in a spoken utterance roughly can be seen as the rhythm of speech. Although our measure is based directly on the speech signal (therefore no recognition system has to be used in advance and no explicit reference to lexical units is required) it is implicitly

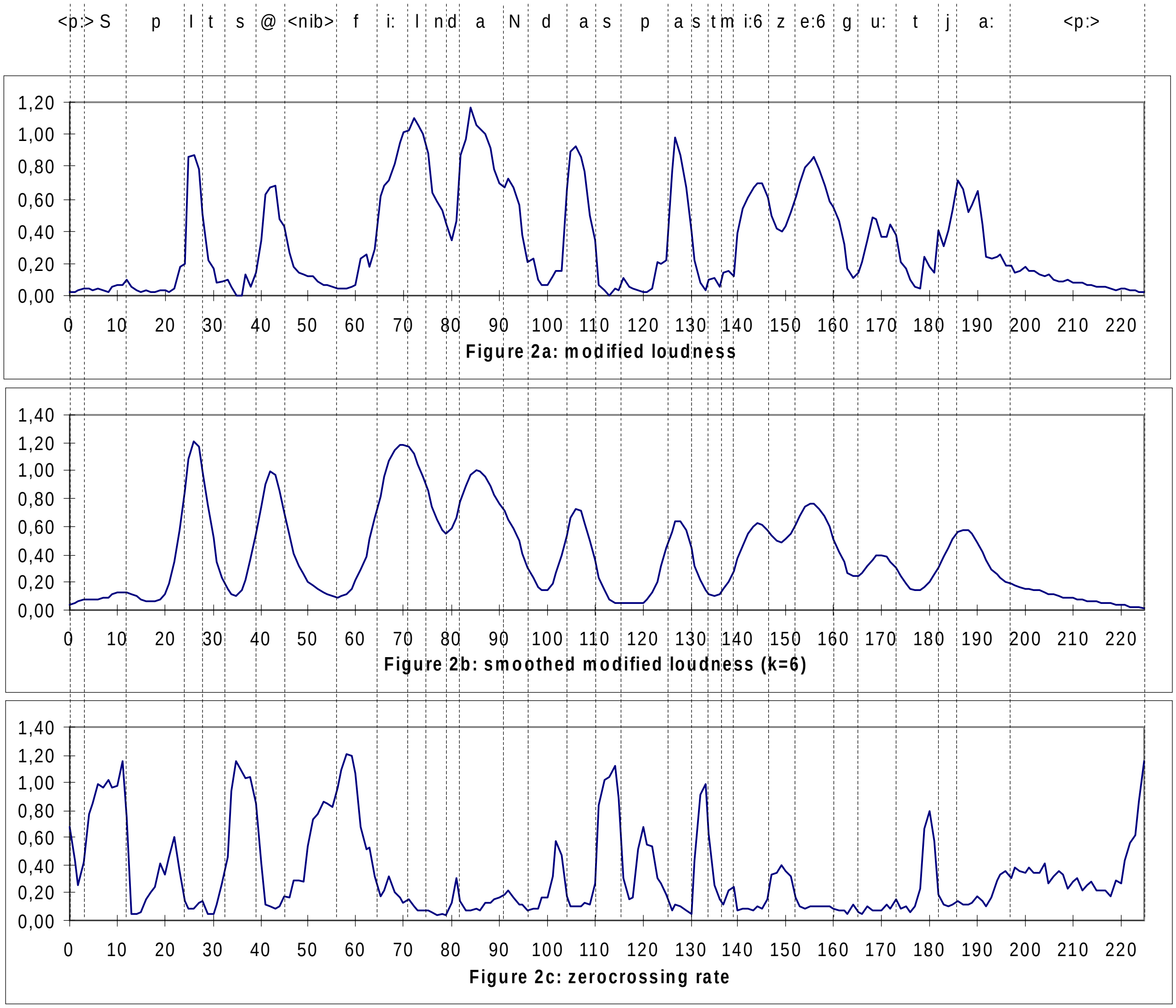

Figure 2: Modified loudness (a), smoothed modified loudness (b), zerocrossing rate (c) and manual transcription of the sentence g073a011 from Verbmobil CD1

linked to a lexical unit, the syllable. Thus it is comparable to lexically-based measures of the speech rate like the phone rate.

It is known from psychoacoustic experiments, that vowels in the syllable nuclei are perceived „louder“ than the neighbouring consonants. A suitable measure for this effect is the modified loudness [7,8]. In contrast to the overall-loudness, it takes into consideration that the main part of the energy of a vowel concentrates on low frequencies whereas for the most consonants the main part of the energy is located at higher frequencies. The modified loudness $N_m(t)$ is calculated as a difference D(t) of the partial loudness functions $N_u(t)$ and $N_o(t)$ (with ν measured in Bark):

$$D(t) = N_u(t) - N_o(t) = \sum_{\upsilon=3}^{15} N_\upsilon(t) - \sum_{\upsilon=20}^{22} N_\upsilon(t)$$

$$N_m(t) = \begin{cases} D(t), \text{if } D(t) > 0 \\ 0, \text{if } D(t) \leq 0 \end{cases}$$

Each specific loudness function $N_\nu(t)$ is defined on the Bark scale giving the loudness within one critical band ν. The modified loudness is computed for every frame (every 10ms) in our standard acoustic preprocessing unit according to [6]. Figure 2a shows the modified loudness calculated by the acoustic preprocessing unit of our standard HMM speech recognition system [9] for each frame of the utterance ‘*Spitze, vielen Dank das paßt mir sehr gut, ja*’. The values of the modified loudness in vowel regions are clearly higher than in consonant regions, thus a detection of maxima has to be done in order to detect vowels.

3. DETECTION OF VOWELS

As one can see in figure 2a not every peak of the modified loudness corresponds to a vowel. For detecting vowels the modified loudness has to be smoothed over time in order to get a kind of envelope of the modified loudness (see figure 2b).

This approach is comparable to [1] where the energy envelope is determined. According to [7,8] a suitable lowpass can be built by a series of k lowpass functions with rectangular impulse response. Increasing k leads to a gaussian impulse response (see figure 3).

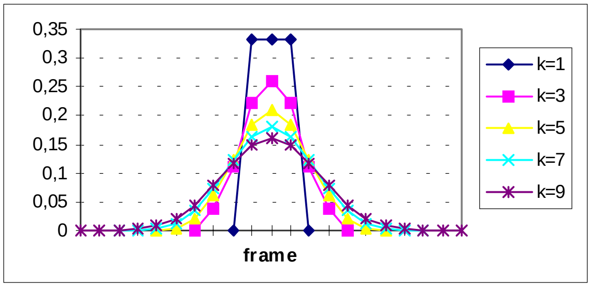

Figure 3: Impulse response of the lowpass for smoothing the modified loudness

3.1. Optimal smoothing function

In the first experiment our aim was to optimize the lowpass function in order to achieve good vowel detection results. As reference for our maximum detector we used a manually transcribed subset of the German Verbmobil database. This subset was divided half and half into a training set and a test set with 694/693 sentences and 17040/18539 vowels respectively. As some syllable nuclei consist of vowel clusters instead of single vowels an inventory of vowels, diphthongs and vowel clusters is defined which build the references for our detector (see table 1).

a:, e:, i:, o:, u:, E:, 2:, y:, a, I, O, U, E, 9, Y, @, 6,
aI, aU, OY,
a:6, e:6, i:6, o:6, u:6, E:6, 2:6, y:6, a6, I6, O6, U6, E6, 96, Y6

Table 1: Inventory of vowels, diphthongs and vowel clusters to be detected (in SAMPA)

Two different measures are used to evaluate our maximum detector. First the vowel error rate (**VER**) is defined:

$$\mathbf{VER} = \left(1 - \frac{\mathbf{hits} - \mathbf{insertions}}{\mathbf{vowels}}\right) \cdot 100\ [\%],$$

where **hits** is the number of maxima, which match to a vowel, **insertions** is the number of maxima, which do not correspond to a vowel in the reference and **vowels** is the number of vowels (and vowel clusters), which can be found in the reference segmentations. As a second measure the correlation coefficient between speech rates calculated from the reference segmentations and speech rates derived from the maximum detector are determined. The two rates are calculated for each utterance using the following definition:

$$\mathbf{ROS} = \frac{\mathbf{n}}{\mathbf{d}}\ [\text{vowels/sec}],$$

where **n** is the number of vowels (detected or given in the reference) and **d** is the duration of the utterance in seconds.

In some preliminary experiments we found out that the steepness of the maxima is an important feature for the indication of insertions, i.e. maxima which do not correspond to a vowel. In general only distinct maxima consistently correspond to vowels whereas flat maxima can often be found in silence regions. For this reason we introduced two additional parameters. The first one is the threshold t, below which the smoothed modified loudness has to fall within a defined frame range d (second parameter) on at least one side of the maximum (see figure 4). This threshold is given as a percentage of the value of the maximum.

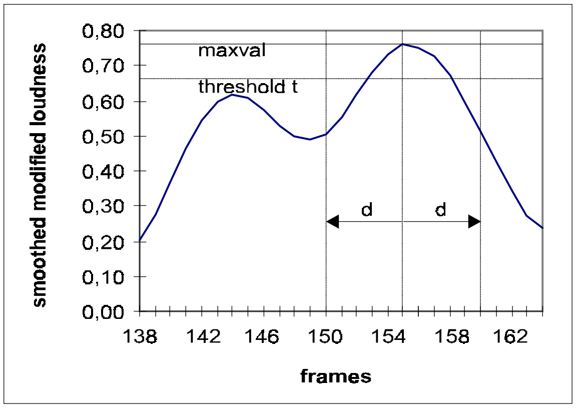

Figure 4: Maximum detection: the smoothed modified loudness must fall below the threshold t on at least one side within the range d

In the last but one line of table 2 (threshold: 'none') error rates without using the threshold t are shown. Using the threshold (table 2: t=0,73...0,93) results in significantly lower error rates. A relative improvement between 8% and 18% can be achieved for the different smoothing functions (k=4...9). The best smoothing with a VER of about 25% can be found for k=6, k=7 or k=8. The corresponding cutoff frequencies of the lowpass functions are 9,67 Hz, 8,98 Hz or 8,43 Hz respectively. This fits quite good to the average syllable rate of 4,5 syllables/sec.

3.2. Additional feature for better detection

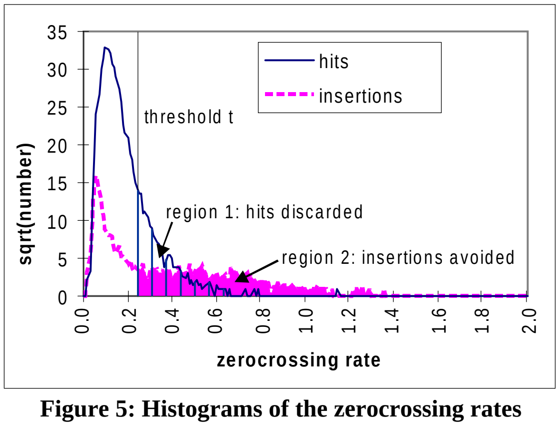

Figure 5: Histograms of the zerocrossing rates of hits and insertions

A closer look at figure 2a-c reveals that the zerocrossing rate can be used to further reduce the vowel error rate. Maxima within regions in which the zerocrossing rate is higher than a defined threshold should be discarded in order to further reduce insertion errors. The threshold to be used can be determined on the training set. Therefore a histogram of the values of the zerocrossing rate has to be built for both correctly detected vowels (hits) and insertions. The optimal threshold can be determined by evaluating the difference between the number of additional errors, which occur by discarding correctly detected maxima (figure 5: region 1) and the number of errors, which can be avoided by discarding insertions (figure 5: region 2). This (negative) difference has to be minimized in order to improve the VER. Resulting error rates on the training set can be seen in table 2 (rows with use of the zerocrossing rate are marked with a ‘zc’). A further improvement of 5% to 8% relatively can be achieved using the optimized threshold for the different smoothing functions.

	smoothing factor k					
peak threshold t	4	5	6	7	8	9
0,73	26,78	25,65				
0,73 zc	24,61	23,83				
0,75	26,97	25,6	25,13			
0,75 zc	24,51	23,68	23,6			
0,77	27,10	25,7	25,14	25,01	25,12	25,65
0,77 zc	24,5	23,66	23,51	23,63	24,03	24,67
0,79	27,29	25,83	25,11	24,98	25,13	25,52
0,79 zc	24,57	23,62	23,37	23,51	23,97	24,5
0,81	27,65	25,97	25,28	25,00	25,07	25,42
0,81 zc	24,63	23,6	23,41	23,45	23,84	24,35
0,83	27,99	26,26	25,48	25,11	25,19	25,42
0,83 zc	24,81	23,67	23,43	23,43	23,81	24,25
0,85		26,53	25,66	25,28	25,33	25,59
0,85 zc		23,8	23,44	23,47	23,81	24,27
0,87			26,02	25,55	25,56	25,65
0,87 zc			23,59	23,54	23,86	24,18
0,89				25,82	25,76	25,82
0,89 zc				23,57	23,85	24,12
0,91				26,29	26,07	26,00
0,91 zc				23,73	23,95	24,15
0,93				26,87	26,55	26,46
0,93 zc				23,96	24,1	24,31
none	32,69	30,21	28,87	28,1	27,82	27,69
none zc	26,97	25,36	24,67	24,42	24,58	24,75

Table 2: VER on the training set for different peak thresholds with and without using the zerocrossing rate

smoothing factor k	peak threshold t	optimized threshold zc for zc-rate	error rate	correlation coefficient
4	0,77	0,42	23,75	0,781
5	0,81	0,42	23,26	0,792
6	0,79	0,42	22,72	0,796
7	0,83	0,44	22,73	0,792
8	0,85	0,47	23,28	0,787
9	0,89	0,43	23,7	0,778

Table 3: VER and correlation coefficients on the test set with parameters optimized on the training set

Finally in table 3 the vowel error rates for the test set (with use of the zerocrossing rate) are shown. The best smoothing factors can be found as k=6 and k=7, which are the same as on the training set. Additionally for k=6 the correlation coefficient reaches a maximum, too.

4. CONCLUSION

In this paper we have presented a new feature-based method for estimating the speaking rate on the acoustic signal. The smoothed modified loudness is used to detect vowels, and the relation between vowels and syllable nuclei enables us to compare our measure to the syllable rate, which is a lexically-based measure. In experiments on the German Verbmobil spontaneous speech database we achieved a vowel error rate of 22,72% on the defined test set. This result is quite encouraging since the correlation to the actual speaking rate, derived from manually transcribed data, is rather high (up to 0,796). Thus the accuracy of our measurement is sufficient to use it for ROS adaptation within the recognition module of our system.

5. ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry for Research and Technology (BMBF) in the framework of the Verbmobil Project.

6. REFERENCES

[1] Morgan N., Fosler E., Mirghafori N., *Speech Recognition Using On-line Estimation of Speaking Rate*, EUROSPEECH ’97, pp. 2079-2082, Rhodes.

[2] Mirghafori N., Fosler E., Morgan N., *Towards Robustness to Fast Speech in ASR*, ICASSP ’96, Vol.1, pp. 335-338, Atlanta.

[3] Mirghafori N., Fosler E., Morgan N., *Fast Speakers in Large Vocabulary Continuous Speech Recognition: Analysis & Antidotes*, EUROSPEECH ’95, pp. 491-494, Madrid.

[4] Siegler M.A., Stern R.M., *On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems*, ICASSP ’95, pp. 612-615, Detroit.

[5] Verhasselt J.P., Martens J.-P., *A Fast and reliable Rate of Speech Detector*, ICSLP ’96, pp. 2258-2261, Philadelphia.

[6] Ruske G., Beham M., *Gehörbezogene automatische Spracherkennung.* In: „Sprachliche Mensch-Maschine-Kommunikation“, (H. Mangold, Hrsg.). Oldenbourg-Verlag, München Wien, 1992, 33-47.

[7] Ruske G., *Automatische Spracherkennung: Methoden der Klassifikation und Merkmalsextraktion*, pp. 117-123, Oldenbourg, München (1994).

[8] Weigel W., Ruske G., *Continuous speech recognition using syllabic segmentation and demisyllable Hidden Markov Models*, EUROSPEECH ’89, Vol.1, pp. 17-20, Paris.

[9] Plannerer B., Einsele T., Beham M., Ruske G., *A continuous speech recognition system integrating additional acoustic knowledge sources in a data-driven beam search algorithm*, ICSLP’94, pp. 17-20, Yokohama.