Automatic measurement of speech rate in spoken Dutch

de Jong, N.H.; Wempe, T.

Publication date
2007

Published in
ACLC Working Papers

Automatic measurement of speech rate in spoken Dutch[*]

Nivja H. de Jong
University of Amsterdam

Ton Wempe
University of Amsterdam

In this paper, we describe a method to automatically measure speech rate without the need of a transcription. A script written in the software program PRAAT detects syllables in running speech. Peaks in intensity (dB) that are preceded by dips in intensity are considered as potential syllable nuclei. The script subsequently deletes peaks that are not voiced. Testing the resulting syllable counts of this script on two corpora of spoken Dutch, we obtained high correlations between speech rate calculated from human syllable counts and from automatically determined syllable counts. We conclude that a syllable count measured in this automatic fashion, suffices to reliably assess and compare speech rates between participants and tasks.

1 Introduction

Becoming fluent in a second language is one of the most difficult aspects of learning a second language (J. H. A. L. De Jong & Van Ginkel, 1992). At the same time, measures of fluency are an important aspect of second language speaking proficiency (see, e.g., the Common European Framework of Reference for Languages (2001), p. 28 – 29). Tavakoli and Skehan (2005) have suggested that three different aspects of fluency can be distinguished: breakdown fluency (number and length of pauses), speed and density per time unit (speech rate), and repair fluency (false starts and repetitions).

For most speaking tests, fluency is a score awarded by human raters who presumably use all aspects of fluency in their judgement. Correlations between

[*] This research was carried out at the Amsterdam Center for Language and Communication. Research funded by NWO under grant number 575-21-009 awarded to Prof. J.H. Hulstijn and Dr. R. Schoonen ("Unraveling second language proficiency"). We wish to thank Renske Berns for her help in counting syllables in a subset of the WiSP-corpus, and Jan Hulstijn, Rob Schoonen, Rob van Son and Paul Boersma for their helpful comments on earlier versions of this paper.

ACLC Working Papers 2:2, 2007: 51 –60

such subjective measures of fluency and objective measures of fluency have shown that *speech rate* is the best predictor of subjective fluency (Cucchiarini, Strik, & Boves, 2002). Kormos and Dénes (2004) show that speech rate in terms of number of syllables per time unit is a good predictor of subjective fluency. In order to ensure objective scores on speaking tests, objective measures of fluency would be preferred. ‘Breakdown fluency’ can objectively be measured by measuring the duration and number of silences in running speech, ‘repair fluency’ can be determined objectively by counting false starts and repetitions, and speech rate, finally, can objectively be measured by counting syllables. However, counting syllables is a tedious job and is often cast aside due to time constraints. In the context of a large-scale research project on the correlates of speaking proficiency carried out at the University of Amsterdam (What is Speaking Proficiency: http://www.hum.uva.nl/wisp), we developed two tools to measure fluency automatically. We wrote a script in the software program PRAAT to automatically detect silence in speech (a simplified version of which is now incorporated in the button *To TextGrid (silences)*. This paper concerns another script in PRAAT that automatically detects syllable nuclei to compute speech rate in terms of syllables per time unit.

Besides for second language research that (wishes to) include a measure of fluency, speech rate is a very important signal for automatic speech recognition as well. Human listeners are able to understand both fast and slow speech in an automatic way. Speech recognizers implemented in computers, however, perform relatively poorly when speech rate is very fast or very slow. In order to improve computer performance, several researchers have proposed that measuring speech rate prior to speech recognition will result in higher success rates of automatic speech recognizers (Pfau, Faltlhauser, & Ruske, 2000) and several ways to automatically measure speech rate in terms of phones and/or syllables per time unit have been put forward.

Mermelstein (1975) developed an algorithm to segment speech into syllables by finding minima in loudness that serve as possible syllable boundaries. Verhasselt and Martens (1996) presented an automatic speech detector that measures phone boundaries and thus calculates rate of speech as phone rate. The phone boundaries are provided by a Multi Layer Perceptron that is trained on a subset of the data that must be hand-segmented at the phone level. Pfitzinger (1999) uses a combination of syllable rate and phone rate to correlate with perceptual speech rate. Syllable rate is calculated by counting peaks in the energy contour, while phone rate is calculated by use of transcription. The syllable, phone, and perceptual speech rates were measured over (very) short stimuli (625 ms). Hunt (1993) used recurrent neural networks to detect syllables. Pfau and Ruske (1998) determined syllable nuclei by detecting vowels on smoothed modified loudness and then calculated speech rate.

All of these different automatic ways to measure syllable and/or phone rate are quite successful. It is difficult, however, to compare the success of these automatic measurers, because they were all used on different corpora, and their success was reported in different ways. Some researchers report correlations between human and automatic speech rate, others report a percentage of syllables (or phones) undetected and falsely detected as compared to human measured syllables (or phones), and yet others report a correlation between the number of manually measured and automatically measured syllables (or phones). The difference in corpora used should also be noted, as some studies used large corpora with many different speakers and others used quite small corpora with few speakers; some used corpora of speech read aloud while others used (semi-) spontaneous-speech corpora. Finally, a noteworthy difference between studies concerns the length of the spurt on which speech rate was calculated. Some studies used extremely short time-windows to calculate speech rate, and others used much longer windows. Perhaps the most obvious reason we cannot compare success of these different automatic speech rate measurers is that the length of the time-window (or spurt) as well as the variance in spurt length will strongly influence calculations of speech rate.

Many of the proposed speech rate measurers need to be trained on a subset of the data that is transcribed or preprocessed by hand (Hunt, 1993; Pfau & Ruske, 1998; Pfitzinger, 1999; Verhasselt & Martens, 1996). In this paper, we will present an easy way to automatically measure speech rate without the use of preprocessing or the need for transcriptions and test it on two different corpora of spontaneously spoken Dutch. To be able to compare the success of the script over these two different corpora, spurt length was controlled. We wrote a script in PRAAT (Boersma & Weenink, 2007) using a combination of intensity (similar to Pfitzinger, 1999) and voicedness (similar to Pfau and Ruske, 1998) to find syllable nuclei.

2 The algorithm

In what follows, we describe the subsequent actions the script completes to find syllable nuclei using intensity (dB) and voicedness. Before the script is run, sound files that are quite noisy should be filtered so that the frequency range is speech-band limited.

Step 1. We extract the intensity, with the parameter ‘minimum Pitch’ set to 50 Hz and using autocorrelation, hence using a window size of 64 ms, using time steps of 16 ms.

Step 2. We consider all peaks above a certain threshold in intensity as potential syllables. We set the threshold to 0 or 2 dB above the median intensity measured over the total sound file (0 dB if the sound is not filtered, 2 dB if the sound is

filtered). We use the median, rather than the mean, to calculate the threshold in order to avoid including extreme peaks in the calculation of the threshold.
Step 3. We inspect the preceding dip in intensity and only consider a peak with a preceding dip of at least 2 or 4 dB with respect to the current peak as a potential syllable (2 dB if the sound is not filtered, 4 dB if the sound is filtered).
Step 4. We extract the Pitch contour, this time using a window size of 100 ms and 20 ms time steps and exclude all peaks that are unvoiced.
Step 5. The remaining peaks are considered syllable nuclei and are saved in a TextGrid (point tier).

The script is available on the personal webpage of the first author (De Jong & Wempe, 2007).

Figure 1 shows a part of a sound file together with the output as TextGrid made by the script. The speech utterance depicted here is *dat uh was wel goed bevallen toen* (‘that uhm was quite well liked then’), totaling 9 syllables, including ‘uh’.

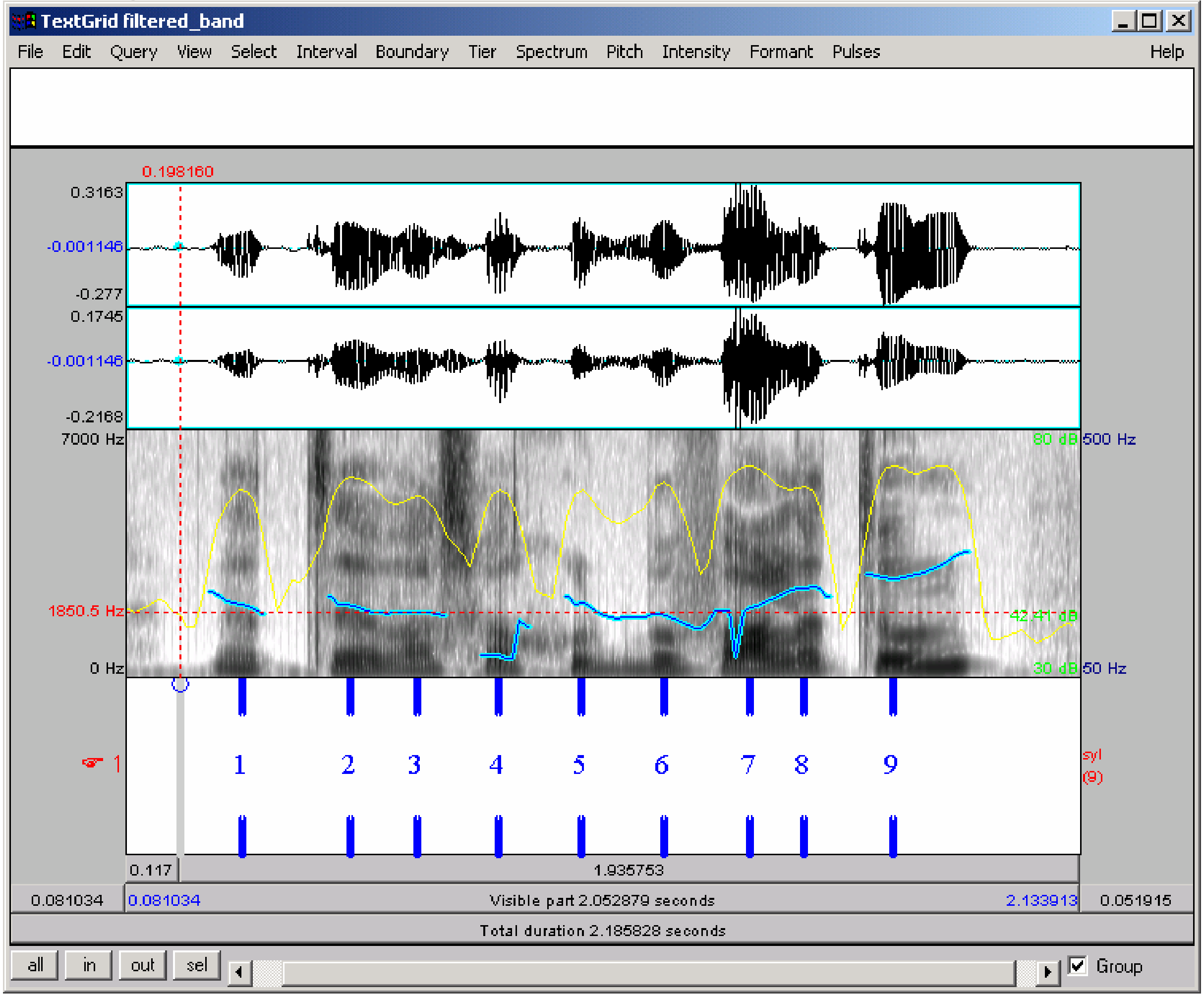

Figure 1: part of a speech file in PRAAT with intensity and pitch shown. The points in the tier are the syllable nuclei as detected by the script

3 Validation

As a part of the project "What is Speaking Proficiency" (WiSP), conducted at the Amsterdam Center for Language and Communication (ACLC) at the University of Amsterdam, we have collected speech data of 258 participants, 200 non-native speakers of Dutch (with various L1s) and 58 native speakers of Dutch. Each participant performed 8 speaking tasks, resulting in a total of approximately 46 hours of speech. See De Jong, Steinel, Florijn, Schoonen, and Hulstijn (in press) for a description of the speaking tasks and an application of the fluency measures. In order to be able to include measures of fluency in our research, we made two scripts written in PRAAT. The first script automatically detects pauses (a modified version is now incorporated in PRAAT in the *TextGrid (to silences*) button), and the second script automatically detects syllables. The second script is described in this paper. In what follows, we report a validation of the computation of syllables per time unit as generated by the script in two different corpora. First, we randomly selected 50 out of the total of 258 * 8 speaking tasks and measured syllables by hand. This corpus comprised 75 minutes of speech. Secondly, we tested the script on a subset of the IFA-corpus that was comparable to the speaking tasks in the WiSP-study (Van Son, Binnenpoorte, Van den Heuvel, & Pols, 2001). This part of the corpus comprised 125 minutes of speech summed over 8 participants.

3.1 Speech data of the Wisp-study

We counted the syllables of fifty speech files. Pauses longer than 0.4 s were considered possible spurt boundaries. We used all spurts of 5 seconds or more to calculate speech rate, and combined consecutive shorter spurts to get to 5 seconds (excluding pauses). We thus avoided calculating speech rate over very short periods of time. We then automatically detected syllables using the PRAAT-script. Many sound files in this corpus were moderately noisy, therefore we filtered all sounds prior to the syllable measuring, using 100 Hz as the lower edge of the pass band, 5000 Hz as the upper edge of the pass band and 50 Hz as the width of the smoothing region. Measuring peaks in intensity (dB), we used 2 dB above the median intensity per sound file as threshold, and 4 dB as minimum dip between peaks, excluding peaks that are unvoiced.

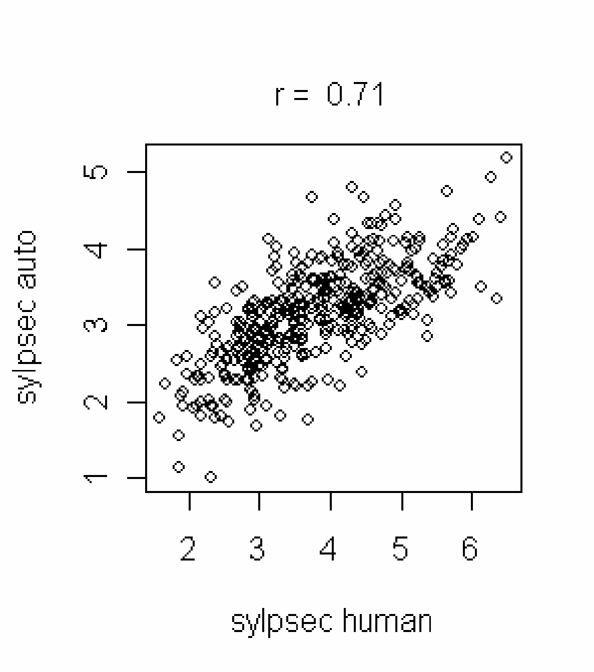

Figure 2: Scatterplot of WiSP-speech data of 50 participants, 441 spurts. Number of syllables per second is calculated per spurt by hand and automatically.

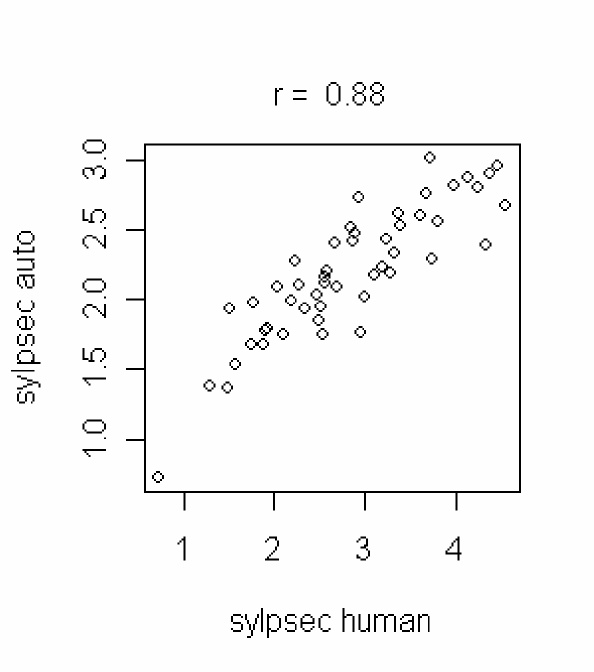

Figure 3: Scatterplot of WiSP-speech data of 50 participants. Number of syllables per second is calculated per task (participant) by hand and automatically.

Figure 2 shows the scatterplot of the human and automatic speech rate calculations per spurt; the correlation was .71. For our purposes of comparing speakers and/or tasks, however, we needed a less refined calculation of speech rate. Figure 3 shows the scatterplot when we calculated speech rate over the total speech file: total number of syllables per speech file divided by total speaking time (correlation .88). In other words, the automatically measured speech rate correlates well with human measured speech rate. However, with these data and these parameters, it seems to be the case that the script tends to miss syllables that were actually present. Upon inspection of the TextGrids made by the script, we concluded that the script misses mostly unstressed syllables that were detected by hand.

3.2 Speech data of the IFA-corpus

The IFA-corpus is an open-source database of hand-segmented Dutch speech. Eight participants (4 female, 4 male) performed several speech tasks, ranging from reading aloud lists of syllables to informal story telling. To validate the script on another corpus of Dutch, we selected the three tasks that were similar to the tasks used in the WiSP-study, eliciting (semi-) spontaneous speech. The three tasks were: informal story telling face-to-face to an "interviewer", retelling of the story previously told, and retelling of a story previously read (Van Son et al., 2001).

For this corpus, we decided not to use a filter, because filtering long sounds takes a lot of time and uses up much computer memory (too much for the computer this script was run on at the time), and because the speech data of this corpus were not as noisy as the above described speech data. As a result, we decided to lower the threshold and minimum preceding dip in intensity. We used the median intensity per sound file as threshold, and 2 dB as minimum preceding dip in intensity. In this corpus, sentences are defined on the basis of pauses as well as syntax, and number of hand-measured syllables could therefore be counted per sentence. As sentences were also defined on syntax, many sentences were very short. Such sentences comprised a single word like "uh", or "en", ("uh" or "and") in mostly beginnings of unfinished sentences.

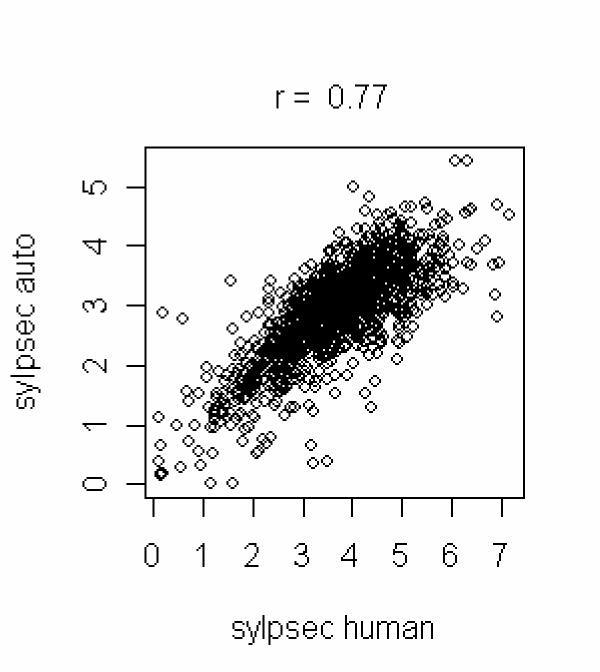

Figure 4: Scatterplot of the IFA-corpus, 8 participants, 1171 spurts. Speech rate, number of syllables per second, counted per spurt by hand and automatically.

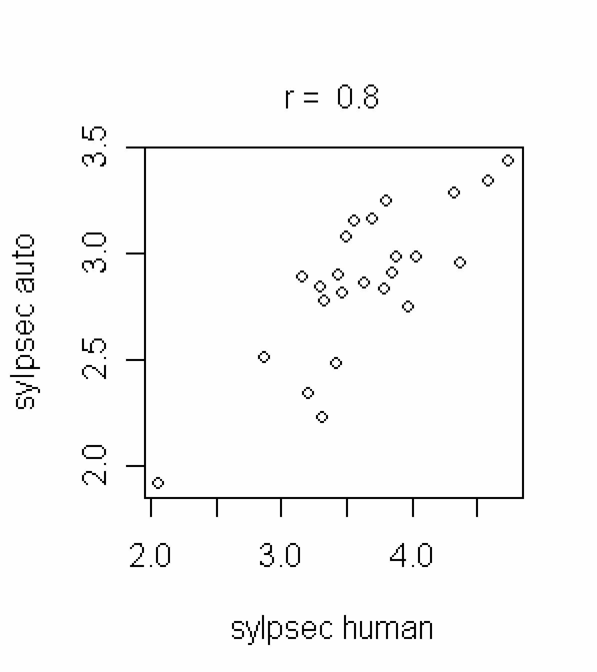

Figure 5: Scatterplot of the IFA corpus, 8 participants in 3 tasks. Speech rate, number of syllables per second, calculated per task by hand and automatically. N = 24.

To test the automatic measures against these existing human-made measures, and to be able to compare success of speech rate measures across the two corpora, we redefined spurts in this corpus as stretches of speech (including pauses) of at least 5 seconds (except when the end of the speech file was reached, in which case the remaining shorter spurt was selected). We then counted the number of syllables using the human transcripts and counted the number of syllables measured automatically for the same time period. In this corpus, we have 8 participants for which human measured information is available in speech tasks quite comparable to those of the WiSP-study. In Figure 4, we show, for all 8 participants, the scatterplot of human measured speech rate per spurt with automatically measured speech rate for that same spurt.

Again, for the purpose of comparing tasks and speakers, we need a calculation of speech rate computed per task. Figure 5 shows the correlation of the 8 speakers in 3 different tasks (r = 0.8). As with the speech data of the WiSP study, the script misses syllables that are detected by hand. An inspection of the TextGrids produced by the script, revealed that most of the undetected syllables were unstressed syllables. We think that many of these unstressed syllables might be phonological syllables and therefore detected when measured by hand, but probably not all are also phonetic syllables in the sense that they are present in the signal in any detectable way. Therefore, we may conclude that the algorithm picks up on prominent syllables. As shown by the correlations between human measures and automatic measures, missing such unprominent syllables does not lead to loss of fit. In other words, although the algorithm cannot find all syllable nuclei, it is able to reliably pick up differences between speech rates.

Research by Kormos and Dénes (2004) suggests that in fact it is the number of stressed syllables that correlates best with subjective fluency. Perhaps it is the case that number of prominent syllables better reflects speech rate in the sense that it measures density of content per time unit. Future research is needed to further explore this thesis.

4 Conclusion

In this paper, we described a script written in PRAAT that automatically detects syllables in sound files of speech. No transcription of the speech data is necessary to run this script. The script takes sound files as input and writes a TextGrid file with syllable nuclei marked in a point tier. In two validation studies, we found high correlations between human measured speech rate and automatically measured speech rate. Although the script misses (mostly unstressed) syllables that are detected by human judges, the correlations suggest that the algorithm works well in predicting the actual number of syllables. We conclude that for the purpose of measuring speech rate as number of syllables per time unit comparing speakers and tasks, this script suffices.

In second language testing (see, e.g., the speaking rubrics of the TOEFL test as reported on the ETS-website (Educational Testing Service, 2004), second language research (e.g., Kormos & Dénes, 2004), as well as in diagnosing different language and speech disorders (Feyereisen, Pillon, & Partz, 1991; Redmond 2004; Shenker, 2006) fluency is an important factor to take into account. The script described and validated in this paper may be useful to easily and objectively measure speech rate in terms of syllables per second without the need to transcribe speech beforehand.

As yet, it is impossible to directly compare the amount of success of the different syllable measurers available. First of all, other syllable measurers have been developed to detect syllables in spoken English or German, which might be different from detecting syllables in Dutch. Furthermore, the different corpora on which the existing syllable measurers have been tested, have been transcribed by different criteria. Finally, researchers report Pearson correlations for speech rate or for number of syllables *per spurt*. However, comparisons are confounded if spurt length is uncontrolled. For longer spurts, a count of one or two extra or fewer syllables will not result in a large deviation of the calculated speech rate. For short spurts, a count of a single extra or fewer syllable will result in an enormous difference in the calculated speech rate. Future research should take these mathematical issues into account when comparing different methods that automatically measure speech rate. In the present paper, we opted for choosing at least 5 seconds as a constant spurt length. In this way, we were able to compare success in syllable detection across corpora.

5 List of references

Boersma, Paul and David Weenink (2007). Praat, http://www.praat.org (Version 4.5.25).

Cucchiarini, Catia., Helmer Strik, and Lou Boves (2002). Quantitative Assessment of Second Language Learners' Fluency: Comparisons between Read and Spontaneous Speech. *The Journal of the Acoustical Society of America, 111*(6), 2862-2873.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

De Jong, John H. A. L. and Lieneke W. Van Ginkel (1992). Dimensions in oral foreign language proficiency. In Verhoeven, Ludo T. and John H. A. L. De Jong (eds.), *The construct of language proficiency: Applications of psychological models to language assessment.* Amsterdam: John Benjamins, 187-205.

De Jong, Nivja H., Margarita P. Steinel, Arjen F. Florijn, Rob Schoonen, and Jan H. Hulstijn (2007). The effect of task complexity on fluency and functional adequacy of speaking performance. In Van Daele, Siska, Alex Housen, Michel Pierrard, Folkert Kuiken and Ineke Vedder (eds.), *Complexity, Accuracy and Fluency in second language Use, Learning and Teaching.* Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunsten, 53 – 63.

De Jong, Nivja H. and Ton Wempe (2007). *Praat script speech rate*. Retrieved December 10th, 2007, from http://home.medewerker.uva.nl/n.h.dejong/bestanden/praatscriptspeechrate.html

Educational Testing Service (2004). *iBT/Next Generation TOEFL Test. Independent Speaking Rubrics*. Retrieved Decmber 10th, 2007, from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Feyereisen, Pierre, Agnesa Pillon, and Marie-Pierre de Partz (1991). On the Measures of Fluency in the Assessment of Spontaneous Speech Production by Aphasic Subjects. *Aphasiology, 5*(1), 1 - 21.

Hunt, Andrew (1993). Recurrent neural networks for syllabification. *Speech communication, 13,* 323 - 332.

Kormos, Judit and Mariann Dénes (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145 - 164.

Mermelstein, Paul (1975). Automatic segmentation of speech into syllabic units. *J. Acoust. Soc. Am., 58*(4), 880 -883.

Pfau, Thilo, Robert Faltlhauser, and Günther Ruske (2000). A Combination of Speaker Normalization and Speech Rate Normalization for Automatic Speech Recognition. *Proc. of ICSLP 2000, Peking, China, 4*(362-365).

Pfau, Thilo and Günther Ruske (1998). Estimating the speaking rate by vowel detection. *Acoustics, Speech, and Signal Processing (ICASSP 2005 Proceedings), 2*, 945-948.

Pfitzinger, Hartmut R. (1999). Local speech rate perception in German speech. *Proc. of the XIVth Int. Congress of Phonetic Sciences, 2,* 893--896.

Redmond, Sean. (2004). Conversational profiles of children with ADHD, SLI and typical development. *Clinical linguistics & phonetics, 18*(2), 107 - 125.

Tavakoli, Parvaneh and Peter Skehan (2005). Strategic planning, task structure, and performance testing. In Ellis, Rod (ed.), *Planning and Task Performance in a Second Language*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 239 - 276.

Van Son, Rob J. J. H., Diana Binnenpoorte, Henk van den Heuvel, and Louis C.W. Pols (2001). The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database. *EUROSPEECH 2001*, 2051-2054.

Verhasselt, Jan P. and Jean-Pierre Martens (1996). A fast and reliable rate of speech detector. *Spoken Language, (ICSLP 96 Proceedings), 4*, 2258-2261.

Contact information:
Nivja de Jong
nivja.dejong@uva.nl
http://home.medewerker.uva.nl/n.h.dejong