The Automatic Alteration of Rhythm in Synthesized Speech

Abstract: *The current perception of synthesized speech does not match the variability of rhythm observed in natural continuous speech. The program described alters the duration of individual phonemes within a synthetic speech utterance, by manipulating the differences in length between stressed vowels, or the inter-stress interval. Phoneme monitoring was used as an objective measure of this method of automatically altering timing in synthesized speech. A program was created to automatically alter inter-stress timing, A significant effect was seen in improving the reaction time of monitored phonemes through changes of the length of inter-stress intervals by the algorithm developed, which supports the original hypothesis. Suggestions are given for integrating this functionality into a speech synthesis system.*

NICHOLAS WINSLOW
Universiteit Utrecht
August 24, 2006

Motivation

In general, rhythm refers to any time-dependent pattern of varying sound intensity, while the underlying pulse is known as the beat or the tempo (Fraisse 1982). As rhythm is observed only in comparing the relative strength of multiple adjacent segments, it can be described as inherently suprasegmental in nature (Nooteboom 1997). Along with tone and lexical stress, rhythm forms what is known as prosody (ibid.). Several acoustic cues are used in determining rhythm, such as a distinct pitch contour and the phenomenon of pitch reset conditioned by the need for respiration (Grabe & Low 2002).

The methodology for measuring rhythm in speech has drawn greatly from the traditions of musical rhythm and poetic meter (London 2004). Within musical theory, an inter-onset interval or IOI is the time between the beginnings or attack-points of successive events or notes, the interval between onsets, not including the duration of the events (ibid). For example, two sixteenth notes separated by dotted eighth rest would have the same inter-onset interval as between a quarter note and a sixteenth note (ibid). This inter-onset interval concept is useful for determining analogous rhythmical units in speech (Patel & Daniele 2003).

Fluent speech is inherently discontinuous (Kelso 1995). The necessity of pauses and acceleration/deceleration intervals characteristic of phrasal and utterance boundaries produces a speech signal which continually varies in acoustics (Zellner-Keller & Keller 2005). The unpredictability of segmental boundaries has lead researchers to postulate isochronic markers as attractors within for

speech production (Port, Tajima & Cummins 1999).

However, neither stresses within stress-timed languages, nor syllables within syllable-timed languages show the expected isochronic regularity (Dauer 1983; Couper-Kuhlen 1993). Instead of strict isochrony, perception is dependent on the a combination of the frequency and distribution of phonetic cues, each competing for perceptual attention (Large & Jones 1999). Thus languages which are 'stress-timed' show a more consistent inter-stress duration and a more inconsistent syllable duration when compared to a 'syllable-timed' language (Grabe & Low 2002).

(Barbosa 2002) modeled rhythm production in Brazilian Portuguese (BP) as the result of a coupled oscillators. The two oscillators are coordinated with each other and one represents the string of syllables and the other the string of syllabic stress. Barbosa concluded that there is no 'apparent, systemic, duration-related stress shift' in BP, but by manipulating w_0, or the degree of coupling between the syllabic and phrasal stress oscillators, the rhythm class of the language can be shifted between syllable-timed and stress-timed and vice versa, subject to 'gestural' constraints.

Research indicates that two interrelated senses of rhythm combine to form the global perception of speech(Grabe & Low 2002). Fast rhythms with beat intervals less than 330 milliseconds(ms) are perceived as a single continuous noise (ibid). A slower rhythm with a beat length greater than 450ms, each rhythmic unit can be perceived as distinct (ibid). Across a language with a simple syllable structure resulting from phonotactic restrictions, the relative difference between the simplest and the most complex syllables is not that great and it is possible to say any syllable under the lower boundary of 330ms (Dauer 1983). Conversely, a language with a wide disparity between the most simple and most complex syllable allowed, due to placing little to no restrictions on syllabic features like complex onset clusters, will have a slower rhythm (Ramus *et al.* 1999). In both cases, the point most easily perceived is the point of greatest sonic energy, or the onset of the stressed vowels (Keating 2003). Vowels represent a largely unconstricted vocal tract, as well as high levels of vocal cord vibration, making them prominent points within continuous speech (Klatt 1976).

To use a more prosaic analogy, timing based upon the more transient noise has been described as *machine-gun rhythm* (Pike 1945). Japanese is said to have a similar rhythm based on a sub-syllabic unit, known as a mora (Tajima 1998). A similar mechanical analogy for the timing of the slower and more distinct rhythm would be to call it *Morse-code* or telegraph rhythm (Pike 1945).

The difference between stress-timed and syllable-timed languages is not as clear-cut as was once considered (Dauer 1983). This suggests a continuum between the two strict conditions, subject to change synchronically due to the pragmatic context of utterances and diachronically due to larger historical trends within the usage of a language.

For example, Mexican Spanish incorporates a number of Nahuatl words (such as *chile*, *mesquite*, *peyote*, and *quetzal*) as a lexical substratum (Canger 1988). Nahuatl itself is the most widely-spoken group of Native American languages in North America, with an estimated 1.5 million people who speak one or another Nahuatl dialect (Gordon 2006). Many Mexican Spanish words of Nahuatl origin words have some version (*-tl*, *-tli*) of the complex absolutive suffix of Nahuatl (Canger 1988). Nahuatl words contained sounds not found in Castillian and many other varieties of Spanish, such as glottal stop, [ʃ], and [k^w] (ibid). The number of permissible syllable structures in Mexican varieties of Spanish is thus expanded, causing a tendency towards stress-timing similar to American English prosody. Kartunnen 1976 (cf. Malmberg 1966) suggested that in Mexican Spanish, unlike in other Spanish-speaking countries, it is vowels which lose strength, while consonants are fully pronounced. This is said to be from the influence of the consonant-complex Nahuatl language by an abundance of bilingual speakers and toponyms (ibid).

The importance of suprasegmental timing information is nontrivial, considering the number of factors which effect and are affected by it. (Klatt 1976) looked at how the individual segments and pauses contribute to information about the context or content of the whole utterance. Duration is the primary perceptual cue for discriminating vowel length, voicing, stress, focus and phrasal position (ibid.). The effect of stress in producing a higher intensity, longer duration and more conspicuous pitch movement led researchers to investigate the effect of stimulus timing on the perception of speech. Timing regularity has been found to significantly improve spoken word perception, without regard to metrical expectancy.(Quené & Port 2005).

Rhythm is a key component, along with vocal stress and pitch, of the collection of suprasegmental information known as prosody (Nooteboom 1997). The difficulty of synthesizing appropriate prosodic elements has long been recognized as a major flaw within speech synthesis (de Pijper 1999). The lack of variation in timing is due to the most popular methods of constructing synthetic voices (ibid.). In concatenative voice synthesis, utterances are assembled from snippets of sound, typically the

transitions between phones, or diphones. But differences between natural variations in speech and the nature of the automated techniques for segmenting the waveforms sometimes result in audible glitches in the output, detracting from the naturalness of the synthesized speech (Sak *et al*. 2006).

All commercial voice dialog systems use either pre-recorded voice segments or a synthesized voice generated with fixed parameters (Ward & Nakagawa 2004). No dialog system can listen to and adjust its production rate to that of users (ibid.). Thus, if a text-to-speech (TTS) program speaks at a rate too slow or too fast the user will end up frustrated, deepening the perception of a lack in naturalness of the synthetic speech produced.

All spoken information within TTS programs is presented independent of user needs and may be too fast for some users, such as non-native speakers, children, people in noisy environments and the elderly, and too slow for others, causing a loss in user, system and connection time (Rye 1999). This mismatch to the conventions of turn-taking and variability of rhythm observed in everyday speech is a major problem with the perception of synthetic speech, suggesting the need for a voice dialog system which alters speech timing based on user specifications (Ward & Nakagawa 2004). In this project, we wished to determine whether greater timing regularity would make synthetic speech easier to understand.

Project Design

The preliminary review of available speech synthesis and voice editing software showed that automatic alteration of inter-stress timing was a functionality unavailable in all systems. Therefore the intermediate phase of this project involved the development of a software module which could translate simple parameters of inter-stress timing within a TTS system. The final phase involved perceptual testing of the regularized output using an objective reaction time-based measure.

Hypothesis

By altering the timing of synthesized speech to match the more regular rate of inter-stress timing suggested to be present in natural speech (as in Quené & Port 2005), the speech generated will be more accurately perceived than the same speech generated by an unaltered version of the TTS engine. This should be demonstrable both through an objective measure, such as reaction time within a phoneme

monitoring task.

Methods

Conditions

The program illustrated here alters the duration of individual phonemes within a synthetic speech utterance. It uses a measure of speech timing known as an inter-stress interval (or ISI), referring to a period of duration beginning with a stressed vowel and ending at the onset of another stressed vowel.

As rhythm is comparative between adjacent units in terms of their relative strength, the choice of the maximum prosodic unit determines the complexity of possible rhythms. Longer sequences of rhythmical units can show a greater range of variation, allowing them to carry more information than shorter sequences. Lists have a specific prosody distinct from fluent speech, for example. A balance was struck by using single utterances as the top prosodic level rather than the word, phrasal, or multi-utterance level.

Since (Quené & Port 2005) showed that regularity in the timing of inter-stress periods improved perception, we hoped to show the same effect within synthetic speech. Thus it was proposed that the two conditions be an unaltered version of the stimulus and one altered to match a more regular inter-stress timing. Using an accent marker which precedes each stressed vowel we can represent simply the boundaries of these intervals. In order to allow future researchers to replicate results, the program is able to read and write files in a common format which are, in turn, able to be fed directly into the TTS engine.

Rhythmicator

Development

While no existing TTS program allowed for the variation of inter-stress timing efficiently enough for the type of manipulation we needed for experimentation, it was concluded that an existing TTS system could be modified to automatically produce the specifications desired. Although several options were considered, the choice of a native Dutch population as test subjects led us to an entirely Dutch TTS system. We also sought a system which allowed for a variety of controls on vocal stress, pitch contours,

and the selection of individual sound units.

In conjunction with Arthur Dirksen of Fluency, a leading Dutch text-to-speech engine, we have developed a software module known as *Rhythmicator* which allows us to automatically alter the ISIs within an MBROLA phoneme (or .pho) file. This program was developed in the Borland Delphi system and runs under the Windows operating system. This module works in concert with the Fluency system and MBROLA tools, allowing researchers to analyze and alter the timing of a TTS utterance based on user-selected parameters. By indicating the placement of stressed vowels in the file, these points can be used as points of maximum perceptual prominence. The period between the stressed vowels can then be altered in relation to each other so that the individual ISIs are lengthened or compressed in proportion to the entire utterance.

Using an accent marker which precedes each stressed vowel, the boundaries of the inter-stress intervals can be simply represented. By varying the length of these intervals automatically, the effect of different timing structures on speech perception can thus be more easily manipulated by researchers.

Within the system developed for this project, the input file format is the PHO format, originally designed for the MBROLA project. The PHO format is a plaintext file where each line (excluding comments) represents: a phoneme label given in the ASCII-type SAMPA encoding; the duration of that phoneme in milliseconds; and optionally, one or more pitch points, each consisting of two integers, a percentage of the duration of the phoneme representing the insertion point and the value in Hertz. Our PHO format also marks the location of stressed vowels with an accent marker, inserted as a comment line directly before the stressed vowel. Comments are marked by lines which begin with a semicolon(;). Finally, all the discrete values described above are separated by a single whitespace from each other. An example line in a PHO file would be:

U 435 75 700 *

where 'U' is the phoneme label, which has a duration of 435 ms, a pitch point which occurs 75% into the duration of the phoneme at a peak of 700Hz, and is marked as a stressed vowel. As the output will need to be read by the TTS engine, it is also in the PHO format as described above, the only difference being the modified durations.

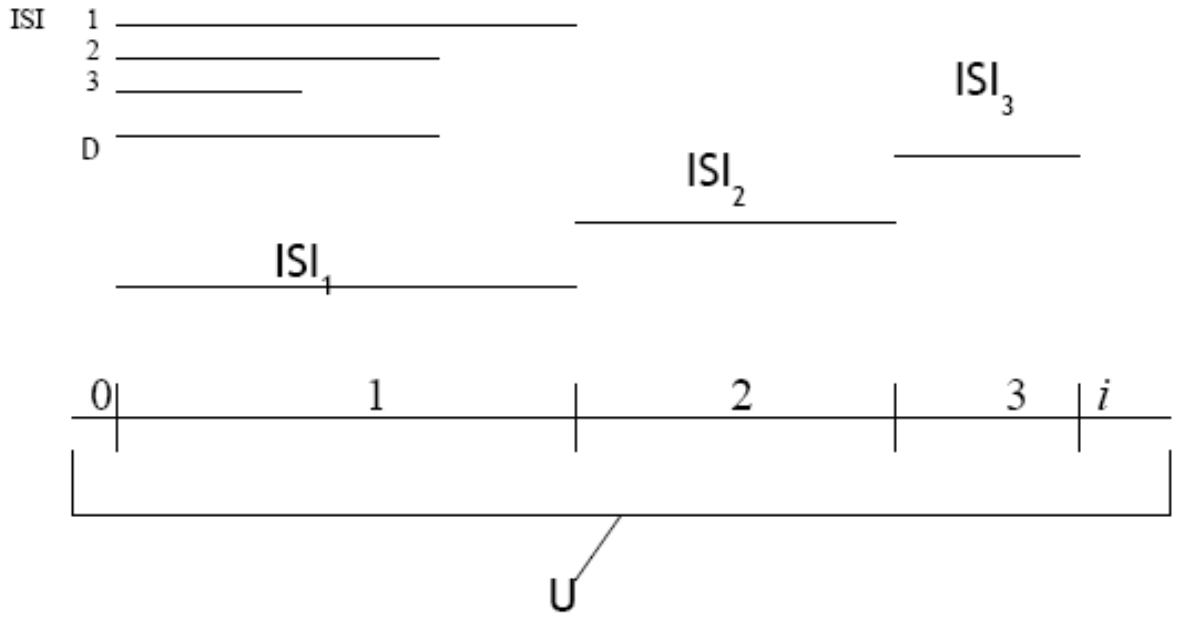
Speech Timing Algorithm

The central component of the *Rhythmicator* program is an algorithm which alters the length of the inter-stress intervals proportionately based upon a set factor set by the program user. Various ad-hoc measures were needed during the operation of the algorithm in order to calculate the correct change in timing to each segment's duration.

Regularity (within the context of this experiment) describes a predefined factor (having a value between 0 and 1) which is used to increase or decrease the difference between the ISIs defined and the average duration calculated. This allows us to make the original duration more or less regular to the ideal. A value of zero indicates a sentence which is not timed or irregular, with no tendency towards a uniform inter-stress interval duration. Whereas, a value of 1 means that an utterance is totally regular in the duration of its inter-stress intervals. For the purposes of this experiment, we have compared unaltered sentences (having a regularity of zero) to sentences which have been altered to be maximally regular whenever possible (or having a regularity of one).

In addition, the original inter-stress intervals were not altered more than 50% of their original length to prevent any jarring alterations in tempo due to the automatic functioning of algorithm. A utterance is fed into the algorithm as a list of segments, grouped into observable inter-stress intervals, and based upon how regular the user wishes to make these ISIs, has a variable added/subtracted from each segment in proportion.

Here is an example to illustrate the functioning of the algorithm. An utterance is fed into the routine which has 3 ISIs: one with a duration of x secs, another with a duration of y secs, and a third with a duration of z secs. The system also disregards the periods immediately before the first stressed vowel and the period after the last stressed vowel. Based on the number of stressed vowels, the system adds up all of the durations and averages them to find a model duration. This model duration is then used in conjunction with an alteration factor (known as regularity, expressed as an integer between 0.0 and 1.0 within this context) to determine the factor for compression/expansion which is applied to each segment to have them come closer to the model duration.



If we we were to represent this algorithm as pseudocode, it would be:

```
Utterances have N ISIs formed by all of the segments S between V*s
Regularity is set at 0 for utterance U (default)
Change Regularity for U to new value (alteration)
        for all segments S in U
                group into ISIs
                divide total duration of U by (number of stressed vowels-1)
                to find D, or the average inter-stress interval
        for all ISI in U and for all segments in ISI
                if S is in ISI AND the duration of ISI is not equal to D then
                        if regularity has a value of 1 go to next
                        else if regularity has a value of 0
                                Calculate the ratio of A over D to
                                determine how much A needs to be altered or the value E
                        else if regularity has a value less than 1 and greater than 0
                                compare E to 1 and use the lesser value as min and the
                                greater value as max
                                subtract min from max and multiply by the difference and
                                add min to get a newly modified version of E
                Multiply segment S by ratio E to give a revised segmental duration
                else go to next S
        else go to next ISI
Return new value for regularity
```

Composition

The main subprograms within *Rhythmicator* include:

- a module which draws and maintains the window on the desktop for the program
- a file chooser with a filter to select PHO files (and only PHO files) to be used
- a file reader which scans the PHO file into a buffered stream
- a tokenizer which parses all of the data

- a module to perform calculations based on the above-described algorithm on the stored data
- a sliding button which allows the setting of a user-defined variable, *regularity*, to be described below
- a file writer which can save a new copy of the PHO file.
- and a direct interface with the TTS software(Fluency), allowing direct playback.

Two topics which we did not have time to expand upon within *Rhythmicator* due to time constraints, but which were notable were: a) whether the system can be expanded across several utterances and, b) how to incorporate half-measures or secondary beats within the system. Concerns about overall stability made limiting *Rhythmicator* to single utterances with a single rhythm the only feasible option.

To test the Rhymicator system, it was necessary to find an objective measurement. Phoneme monitoring allowed us to use reaction time as a standard for ease of perception. Based on prior studies of timing regularity, it was predicted that by altering an utterance in synthetic speech to match a more regular inter-stress timing, the overall utterance would be more easily perceived.

Phoneme Monitoring

Due to the nature of the output as single contained utterances, we needed a method of testing on this level which would allow us to objectively measure the perception of *different* versions of the *same* item. Phoneme monitoring was selected (instead of subjective measures such as user opinion scoring) because it is a proven measure of speech perception within rhythmical experiments and would allow us to form results capable of being generalized with greater reliability (Connine & Titone 1996).

Participants were asked to listen to several sets of sentences in synthesized speech. For each sentence in the experiment, participants were asked to listen for a specific phoneme which may or may not be in that sentence and to press a button on a peripheral button box as soon as they hear it. Each set of sentences was based upon the particular phoneme to be listened for and preceded with a female voice which indicating that phoneme. So for the /t/ set, the instructions would be “Let nu op de 't' als en 'tango'” *(Eng: 'Listen now for the 't' as in 'tango.'”)*

Participants received one of two stimulus lists, each of which has an equal number of control(unaltered in tempo) and variable(altered in tempo by *Rhythmicator*) sentences, representing the two conditions of this experiment. The two lists complement each other so that a control sentence in one list has an

altered version of that sentence in the opposite list, and vice versa.

The reaction time of the participant in detecting the given phoneme is calculated by taking the total reaction time from the beginning of the sentence until a button is pressed and subtracting from this the offset time of the target phoneme within the sentence.

Participants were tested using an audio stimulus presented over headphones and a visual stimulus on either a laptop used for field experiments or within a sound-proof cabin in the University phonetics laboratory. Participants were identified by their first name, gender and their year of birth.

Stimulus Material

Utterances are all in Dutch and are taken from Roeleveld (2004). They are grouped in blocks based upon the target phoneme in each sentence. This provides a well-formed data set for phoneme matching and rhythm. This corpus is used to make two sets using the Fluency text-to-speech engine: a **control** set of unaltered sound files; and a **variable** set with durations altered as a result of the project algorithm.

Each of the two stimulus lists is comprised of 80 test items divided equally between the two conditions, 36 altered filler sentences, 36 unaltered filler sentences and 5 instructional items (natural speech utterances which indicate to the participant the phoneme being monitoring within the sentence) making the total number of items to be 157. The phonemes being monitored in the experiment are [/k/, /p/, /t/, /b/, and /d/]. Experimental and control expressions vary in a pseudo-random order for each subject to cancel out order effects.

Participants

Volunteers were solicited from the general public through the Utrecht Institute of Linguistics-OTS at Universiteit Utrecht to comprise a sample of adult native speakers of Dutch. A total of thirty participants took part in the phoneme matching experiment in the months of July and August 2006. Of this total, 19 were male and 11 were female. They were divided into two groups, each of which received a different member of a pair of matched lists. The lists complemented each other in the items presented, so that an equal number of participants for each list was required. As a result of a number of absent scores in two participants, their scores were eliminated in the final analysis, which used an equal number of 14 participants for each of the two stimulus lists for a total analyzed participant group of 28.

Results

The thirty participants were divided into two groups of fifteen each. Within each group, their response time for individual items was averaged and compared across experimental conditions. Participants were tested during the months of July and August 2006.

Within each test, there were 39 items in the experimental condition (due to one incomplete item), 40 in the control condition, and 40 filler sentences which lack a stimulus, 40 foil sentences which contained a stimulus but were not analyzed, and 5 instructional sentences, making a total of 156 items in each stimulus list.

Combined there were a total of 2212 reaction times elicited. Of these reaction times, 262 (or 12% of the total) had to be discarded because they were either negative values (indicating a premature button press) or false negatives where no button was pressed after a monitored phoneme. Responses were also considered premature if they occurred before 50ms after the target. This left the final number of responses analyzed as 1950. The total time for each test was between 17 and 20 minutes, depending on the time taken by the participant in the practice section.

For each participant, the general range of average reaction times was between 1000ms and 400ms.

Analysis

In a repeated-measures ANOVA within a General Linear Model (GLM), there was a significant decrease in the reaction time between the experimental and control conditions of the test sentences. Between conditions, there was an average reaction time of 610ms for the unaltered version and an average reaction time of 469ms for the altered version.

Thus there is significant evidence to support the original hypothesis stated above *(F(1,14) =8.08; p=.013)*. This suggests that increasing the timing regularity of synthesized speech improves its perception.

Discussion

Our original motivation was to see whether timing regularity could improve the perception of synthetic speech. Although this required an extensive process involving the development of a novel software process, there was no guarantee that a substantial effect would be seen.

Bearing that in mind, there are alternative plausible explanations for the results. The sample population could require a greater diversity in demogrpahics to accurately reflect the larger total population of native Dutch speakers. There could be artifacts in the visual or audio stimulus which confound their perception. The length of the test could have caused a fatigue effect on the later items.

However, several phonetics experts within the department have examined the test corpus and found no observable artifacts. Also the total length of the test was similar to other phoneme monitoring experiments. While a more diverse participant sample would have been ideal, time constraints prevented a wider group from being found for testing.

Since there appears no evidence of these confounding effects in our experiment, we can cautiously claim that our hypothesis has been supported: *synthetic speech which shows timing regularity is* ***more accurately perceived*** *compared to synthetic speech without regular timing of its inter-stress periods.*

This suggests that future TTS systems could improve their connection to natural speech rhythms by incorporating a variant of the system we have described, as more regular timing of inter-stress periods had a positive effect on the overall perception of synthetic speech. Recognizing the need for variable timing would allow for the TTS systems to match the natural differences in speech rhythm used by native speakers. This type of user-adaptive modification could serve to bridge speech recognition and speech synthesis into a model more reflective of the varying conditions of language use.

This project hopes to serve as strategic research towards a more natural text-to-speech system by demonstrating the importance of inter-stress regularity in synthetic speech. Improvement in the perceived naturalness of synthetic speech would help to provide timing customized to the situation and resistant to the 'fatigue' reported by habitual users of systems currently available. Future research could include the incorporation of secondary rhythmical structures, as well as a better feedback to measurements of speech timing from statistical recognition of the user's speech.

LITERATURE

Barbosa, P. (2002). "EXPLAINING Brazilian Portuguese resistance to stress shift with a coupled-oscillator model of speech rhythm" Cad. Est. Ling., *Campinas (43)*, 71-92.

Canger, U. (1988) "Nahuatl dialectology: a survey and some suggestions". *IJAL 54(1)*.

Connine, C. & Titone, D. (1996) "Phoneme Monitoring", *Language and Cognitive Processes 11 (6)*, 635-645.

Couper-Kuhlen, E. (1993). *English Speech Rhythm: Form and Function in Everyday Verbal Interaction,* John Benjamins.

Cummins, F. & Port, R. (1998) "Rhythmic Constraints on Stress-timing in English", J*ournal of Phonetics 26*, 145-171.

Dauer, R.M. (1983). "Stress-timing and Syllable-timing Reanalyzed."*Journal of Phonetics (11)*, 51-62.

Fraisse, P. (1982) *Rhythm and Tempo in the Psychology of Music*, Academic Press.

Fowler, C.A. (1979). "'Perceptual Centers' in Speech Production and Perception." *Perception & Psychophysics 25 (5)*, 375-388.

Gordon, R.(ed.) (2006) *Ethnologue*, SIL, online version.

Grabe, E. & Low, E.L. (2002). "Durational Variability in Speech and the Rhythm Class Hypothesis" *Papers in Laboratory Phonology VII,* Mouton de Gruyter.

Kartunnen, F. (1976) *Nahuatl in the Middle Years*, Los Angeles.

Keating, P. (2003) "Phonetic Encoding of Prosodic Structure" in *Proceedings of the 6th Int'l Seminar on Speech Production*.

Kelso, J. (1995) *Dynamic Patterns: the self-organization of brain and behavior*, MIT Press.

Klatt, D. (1976). "Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence" *JASA 59 (5)*, 1208-1221.

Large, E.W., & Jones, M.R. (1999). The Dynamics of Attending: How People Track Time-varying Events. *Psychological Review 106(1),* 119-159.

London, J. (2004). *Hearing in Time: Psychological Aspects of Musical Meter.*

Morton, J., Marcus, S., Frankish, C. (1976). "Perceptual Centers." *Psychological Review (83)*, 405-408.

Nooteboom, S. (1997) "The Prosody of Speech: Melody and Rhythm" in *the Handbook of Phonetic Sciences*, Blackwell, 640-673.

Patel, A. & Daniele, J. (2003) "An Empirical Comparison of Rhythm in Language and Music",

Cognition 87, B35-B45.

de Pijper, J. (2004) "High-Quality Message-to-Speech Generation in a Practical Application" in *Fundamentals of Speech Synthesis & Speech Recognition*, (ed. E. Keller) Wiley & Sons.

Pike, K. (1945) *The Intonation of American English*, University of Michigan.

Port, R.F., Tajima, K. & Cummins, F. (1999). "Speech and Rhythmic Behavior" in *Non-linear Development Processes.* Royal Dutch Academy of Arts & Sciences.

Ramus, F., Nespor, M. & Mehler, J. (1999) "Correlates of linguistic rhythm in the speech signal", *Cognition 73,* 265-292.

Roeleveld, A. (2004). "Temporele Regelmaat in Verbonden Spraak" MS, Utrecht University.

Rye, J. (1999) "Speech Synthesis at Higher Speaking Rates", *Proceedings of CSUN '99.*

Quene, H. & Port, R.F. (2005). "Effects of Timing Regularity and Metrical Expectancy on Spoken-word Perception." *Phonetica 2005 (62),* 1-13.

Sak, H., Gingor, T., & Safkan, Y. (2006) A Corpus-Based Concatenative Speech Synthesis for Turkish", *Turkish Journal of Electrical Engineering*.

Tajima, K. (1998) "Speech Rhythm in English and Japanese", MS, Indiana University.

Van Santen, J.P.H. & Shih, C. (2000). "Suprasegmental and Segmental Timing: Models in Chinese and American English" *JASA 107 (2)*, 1012-1026.

Ward, N. & Nakagawa, S. (2004) "Automatic User-Adaptive Speaking Rate Selection" *Int'l Journal of Speech Technology (7)*, 259-268.

Zellner-Keller, B. & Keller, E. (2005) "The chaotic nature of speech rhythm," in *Integrating Speech Technology in Language Learning* (ed. Delcloque, P. & Holland, V.), Swets & Zeitlinger.

Appendix: *Test Corpus*

(Roeleveld 2004)

Lijst 1

Mijn zus houdt het meeste van nasi maar ik vind **bami** nog lekkerder.
Mijn opa speelde heel mooi vlooi maar mijn oma wou dat hij **banjo** ging spelen.
Om goed te worden in je studie moet je eerst een **basis** van kennis verweven.
Op de prairie in Amerika zwief de **bizon** in grote kuddes.
Hij zou in de bergen op een wilde rivier met zijn **kano** gaan varen.
De dierenartsvertelde bezorgd aan het baasje dat zijn hond een **kilo** te veel woog.
De auteur van de roman wilde nog een punt en een **komma** wijzigen.
Iedereen weet wie de premier is van Nederland maar wie is de **koning** van Swaziland?
De oude man vindt het fijn om meet een **pelgrim** door het land te wandelen.
De telfoniste trachtte's avonds tijdens het avondeten de student een **polis** aan te smeren.
In de dierentuin van Madrid was de **panda** een ware attractie.
Van alle snoepjes in de winkel wilde de dreumes perse een **toffee** hebben.
De professor kwam een half uur voor de cermonie zonder zijn **toga** binnen gelopen.
De beeldhouwer was bijna klaar maar wilde nog even de **torso** bekijken.
Mijn zoontje van acht maakte een tekening met een **potlood** van zijn broertje.
In Buenos Aires leerde mijn vriendin de **tango** dansen.
Sommige mensen nemen na het avondeten liever geen **koffie** maar thee.
Slechts een kwart van alle mensen heeft er voor gekozen om zich als **donor** te registreren.
Mijn oude buurvrouw houdt een praatje als ze lang voor de **kassa** moet wachten.
Een bruinachtig katdier uit Noord-Amerika wordt een **poema** genoemd.

Lijst 2

De zeer oude pop-muzikant maakte een album na jaen van afwezigheid.
De toerist meende te zien dat de **fakir** zich verwondde.
De fotograaf dacht dat hij voor het portret de juiste **focus** had genomen.
Een onvoldragen mens in de baarmoeder is een **foetus** en geen baby.
Het bewijs werd door middel van een **foto** duidelijk voor iedereen.
Als je iets via internet bestelt kan je het **franco** laten bezorgen.

Als je hard gooit komt de **frisbee** wat verderdan anders.
Aan het einde van de dag ging de reiziger een **herberg** zoeken.
Ondanks z'n afkeer van gzelschap kwam de **hertog** naar de vergadering.
De leraar rende snel naar de klas waar de **nota** van de leerling lag.
Een schip mag niet uitvaren als z'n **radar** kapot is.
De hoofdredacteur slaagde erin om iedere keer een **rebus** te maken.
Alleen als je goed tegen hitte bestand bent kan je in de **rimboe** gaan wonen.
Mijn Indische oma vind het fijn om een maaltijd met **sambal** te maken.
Op het zonnige terras was geen plek zonder **schaduw** te vinden.
Als je een splinter hebt moet je je **vinger** in water met soda stoppen.
De tiener was erg bang dat er een **vampier** in de kast verscholen zat.
Het was in het begin van de negende eeuw dat een hemd werd een **wambuis** genoemd.
In veel restaurants staan mooie bloemen om de gasten **welkom** te heten.

Lijst 3

Na jarenlang zoeken vonden de artsen in Thailand een nieuw **bacil** in de jungle.
De vriendin van mijn moeder is zoekende naar **balans** in haar leven.
Het kind keek verwonderd om hoog toen de **ballon** wegvloog en ging daarna hard huilen.
In de dierentuin zag ik een aap die op een tak een **banaan** zat te eten.
In adelijke kringen is het niet vreemd als een hertog of een **baron** zijn eigen honden meeneemt tijdens de jacht.
Na een vermoeiende dag in de woestijn is het fijn om van je **kameel** af te stappen.
De jongen peinsde zich suf over de vraag waarom een **kanaal** recht is en niet slingert zoals een rivier dat doet.
Onze glazen tafel viel in duizend scherven toen de buurvrouw de **karaf** liet vallen.
Herman vraagt zich altijd af hoe je zonder cement een **kolom** kan bouwen.
Na het horen van zijn piepstem was het een verrassing dat de spreker een **kolos** van een man bleek te zijn.
Iedereen bij mij op school wilde vroeger een leuk schattig huisdier ik wilde een **konijn** en mijn vriendje een goudvis.
Vroeger legde men tuinen aan die geheel moesten aansluiten bij de stijl van het **paleis** dat er naast lag.
In Belgie zijn veel mensen al jaren aan het zoeken naar het **paneel** van de Rechtvaardige

Rechters.
Vermeer was een Hollandse schilder die grote indruk met zijn **penseel** kon maken.
Niet alleen frikadellen zijn typerend voor het Nederlandse eten maar ook een **puree** wordt vaak gemaakt.
Het dwergvolk dat leeft in de afgelegen streken van Afrika heet **pygmee** en niet bosjesman.
Voor het afhandelen van de verzending overzee is een hoger **tarief** bedongen.
De fabriek begroef jarenlang illegaal afval op een verlaten **terrein** aan de rivier.
Een hoog podium waarop je de spelers kan zien, is handig als je **toneel** wil bekijken.
De oude visser is gelukkig zolang hij maar **tonijn** mag vangen.

Lijst 4

Niet iedereen die last heeft van angst hoeft dan aan **fobie** te lijden.
Na een middag piano spelen zijn de kinderen naar de **fontein** gelopen.
Ik vind een mandarijn net zo lekker als een **framboos** of een druif.
Mijn jarige neefje was helemaal blij door de nieuwe **gitaar** van zijn ouders.
De walvis zwom weg van de schutters die snel een **harpoen** wilde afvuren.
Mijn buren hebben in Frankrijk een **hotel** kunnen kopen.
Uit een vragenlijst is gebleken dat Beckham het grootste **idool** van de jongeren is.
De bediende van de koning wordt een **lakei** genoemd.
Ik heb voor mijn verjaardag een bureau van **metaal** gekregen.
Mijn moeder heeft na jaren lang zeuren eindelijk een **mobiel** gekocht.
De tekenaar vroeg zich af wie hij als **model** zou vragen.
De verdediging wilde graag van de aanklager een **motief** voor de inbraak horen.
Een gasbedrijf is boos op een commissie zolang die de **natuur** wil beschermen.
Iedereen weet dat een smaragd meer dan een **robijn** waard is.
De lastige jongen vroeg aan zijn broer om zijn nieuwe **sandaal** te pakken.
De prinses vroeg vriendelijk aan de bediende of hij lakens van **satijn** kon pakken.
De scepter om mee te zwaaien is het ultieme **symbool** van de macht.
Op een raam in de winkelstraat heeft een **vandaal** zijn naam gekrast.
Iedereen vond het ongelofelijk dat de man de **vulkaan** bezocht.
Vier jaar geleden is de moade van de **woestijn** naar de oase gegaan.

Nepzinnen

Toch denkt militair dat de vijand naar hem kijkt.
Ik besloot onder de daklijst van het theehuis te schuilen totdat de bui over was.
Sinds mijn terugkomst in het land overvalt mij een kilheid.
Met het verstrijken van de middag begon mijn hoofd raar te bonzen en voelde ik me misselijk worden.
Vroeger aaide ik uren land de katten van mijn tante maat tegenwoordig aai ik liever mijn eigen kat.
Het licht breekt door de brand geschilderde ramen in lange waaiers van kleurig blauw
Ik leerde mijn eerste Engels uit een oud grammaticaboek van mijn moeder.

Op die middag kwam tijdens de pauze de tovenaar naar ons toe om ons te verrassen.
Aan het einde van de week begon het 's avonds te stormen.
Een peuter moet elk jaar een keer langs de dokter om een prikje te halen.
Ondertussen waren we om het restaurant heen gewandeld en stonden we weer bij de ingang.
In 1613 trachten de Zeeuwen een kolonie te vestigen in Suriname.
Ik ging lekker op de stoel zitten en liet mij door mijn moeder verassen.
Het was droog maar het water liep in kleine straaltjes door de goten.

Ik ga in de andere kamer een kimono voor je klaarleggen.
Ik klom snel langs de ladder naar boven en duwde daar de ijzeren deur open.
De volgende morgen zag ik de kakkerlak snel over de keukenvloer rennen.
De ui kan heel fijn gesneden worden en daarna worden gebakken.
Doe de bloem in een schaal en roer er vier of vijf eieren doorheen.
ik brandde mijn lippen aan de vieze warme koffie die de ober mij gaf.
Mijn opa riep om hulp toen een plank pijnlijk op zijn schouder viel.

Pablo zit op een verroeste bank en heeft prachtig uitzicht op de aanlegsteiger.
Ik bekijk een kralenketting naast de kassa en zie een muis onder een plint wegschlieten.
Ik zoek al jaren naar een verzamelboek waarin ik mijn collectie zilveren munten op kan bergen.
Gisteren lag mijn kat lekker te slapen in het washok op een stapel gestrekken lakens.
De serveerster gaf aan al haar klanten een kopje koffie met een lekkere bonbon.

Lekker warm ingestopt met een kriuk bij haar voelen viel ze snel in slaap.
Nieuwsgierig keek ze naar hem en zag ze een glinstering in zijn ogen.

Het meeste wat hij hier vertelde moest de vrouw toen nog zien.
Hij stond midden in de nacht op om de vissers hun boten in de zee te zien duwen.
Hij had drie dagen om zijn gedachten te ordenen en te beslissen of hij de nieuwe baan zou nemen.
Een ander interessant onderwerp was de juridische positie van de slaaf in zijn moederland.
De wind deed het vuilnis door de verlaten straten opdwarrelen.
Mijn broer en zus waren nooit naar een psycholoog of therapeut geweest.
Het is niet meer dan logisch dat hij op dit moment een beetje in de war is.
De buurvrouw had niet in de gaten dat ze haar sleutels vergat toen ze de deur dicht deed.

Het meisje met de paardenstaart en met het witte shirt lachte vriendelijk naar me.
De tween hartsvriendinnen zaten te proesten van lachen onder het genot van een ijsje.
Het is alweer een maand gelden dat ik overheerlijke pannenkoeken heb gebakken voor mijn vriendin.
De stereo die ze van de buurman had geleend produceerde al meer dan een uur hetzelfde deuntje.
Al uren lag ze met hevige pijn in bed naar de muur te staren.
Het meisje is angstig voor paarden omdat ze vindt dat deze zo groot zijn.
Mijn huisgenote poedelt graag enkele uren in bad nadat ze deze heeft schoongemaakt.

Het meisje had kleine oogjes omdat ze de hele nacht een avontuurlijk boek had gelezen.
Mijn moeder zoekt eerst naar haar bril als ze de krant wil gaan lezen.
Het zojuist getrouwde stel kocht een tweedehands bad wat niet door de deur kon.
Het ondeugende meisje hield ervan om met haar broer kattenkwaad uit te halen.
De jongens vernielde de nieuwe bus omdat ze vonden dat het stoer was.
In de winter is de brug vaak dicht omdat er dan weinig waterverkeer is.
De lekke boot van mijn chgrijnige zwager kan niet meer varen.

De poes sloeg speels met haar klauw een touw weg.
De gierige tante skieg haar man als de energierekening weer eens erg hoog was.

Mijn oma was hier op de thee ondanks de hevige regen.
Binnen enkele tellen viel de dronken man op de vloer.
Drie dagen geleden kwam mijn tante op de koffie.
De koerier ging snel in de tas naar het pakje zoeken.
Over twee maanden gaan duizenden mensen gedwongen verhuizen.

Op en dag zal ik de loterij winnen zei hij al twintig jaar lang.
Het duel heeft alles wat een mooie finale moet hebben spanning en sensatie.
Sinterklaas loopt over daken om zijn pakjes af te leveren.
Hoe kan je makkelijk een moeilijke deling oplossen.
Er zijn veel manieren waarop een doek opgevouwen kan worden.
Veel mensen dromen vaak over vere reizen naar exotische plaatsen.
Het bestaan van demonen is niet bewezen door een onafhankelijke.

Het pas getrouwde koppel wilde niet wachten om op reis te gaan.
De jongen wil nieuwe kippen en gaat deze bij de boer halen.
De meester laat het krijtje steeds op de grond vallen.
Mijn oma borstelde het kleed alitijd stralend schoon met groene zeep.
Mijn vader vindt dat mijn kalende broer later het goedlopende bedrijf moet overnemen.
De boer was de koeien in het weiland aan het verzorgen.
De Belg hield niet van Nederlandse kaas en vroeg aan de vrouw om Franse.
De handige jongen krabbelse al snel over bevroren sloot.