TIMING PATTERNS IN FLUENT AND DISFLUENT SPONTANEOUS SPEECH

Douglas O'Shaughnessy
INRS-Telecommunications, Université du Québec
16 Place du Commerce, Verdun, Quebec H3E 1H6 Canada

ABSTRACT

We examine and model global speaking rate, how it varies for both fluent and disfluent spontaneous speech, in terms of the linguistic content of the utterances. Speakers tend to maintain a fixed speaking rate during most utterances, but often adopt a faster or slower rate, depending on the cognitive load (i.e., slowing down when having to make unanticipated choices, or accelerating when repeating some words). Such a model can find application in automatic speech synthesis and recognition, because most synthesizers maintain a constant (and unnatural) speaking rate and most recognizers are not capable of adapting their templates or probabilistic models to reflect global changes in speaking rate.

1. INTRODUCTION

Most previous acoustic analysis of speech has examined data from speakers who carefully pronounce their speech, usually by reading prepared texts. Natural spontaneous or conversational speech differs from careful or read speech, especially concerning hesitation phenomena and variable speaking rates. In spontaneous speech, people often start talking and then think along the way, with interruptions and large changes in speaking rate [Chafe, 1980]. The specific acoustical phenomena studied here concern the timing patterns of spontaneous speech. In particular, we examine and model global speaking rate, how it varies for both fluent and disfluent spontaneous speech, in terms of the linguistic content of the utterances.

In fluently read speech, most speakers maintain a fairly uniform speaking rate, except for brief periods of slower speech just before major syntactic boundaries (e.g., prepausal lengthening). In spontaneous speech, on the other hand, hesitation pauses and restarts are prevalent, and these phenomena have large effects on speaking rate. Read speech is usually spoken at a faster rate than spontaneous speech (e.g., 3 words/sec vs. 2.5 words/s in one study [Snidecor, 1943]). In restarts (or false starts), which are interruptions in the flow of speech, the speaker (usually after a brief pause) reiterates a portion of the speech immediately preceding, with or without a change. The repetition can range from a portion of a syllable up to several words.

A model of timing patterns should find application in automatic speech synthesis and recognition. A major flaw in virtually all synthesizers concerns the repetitious timing patterns that they produce. Even some of the best synthesizers repeat the same basic intonation pattern from sentence to sentence, resulting in tedious output speech, despite reasonable intra-sentence variations. Most research into improved durational models for synthesis has gone toward modeling duration at relatively low levels (i.e., durations of phonemes and syllables, based on phonetic context and stress patterns).

Most variations in timing patterns, and in duration generally, are not handled well in current recognition systems. Some low-level phonetic duration aspects are directly encoded as geometric probability distributions (pdfs) in phonemic hidden Markov models (HMMs). In word-based HMMs, the durational effects of word-level stress can also be so encoded. However, larger durational variations, e.g., due to speaking rate changes and sentence-level stress, are mostly ignored in HMM systems, which can hinder accurate recognition performance. Continuous-speech recognizers are largely based on phonemic HMMs, whose state transition pdfs are biased toward average phone durations. When disfluencies cause much slower speech, the frame independence assumption of HMMs leads to many spurious phoneme insertions. Thus, a better knowledge of how speaking rate varies in spontaneous speech should be of assistance in designing future speech recognizers that propose to handle conversational speech.

2. PREVIOUS STUDIES

To quantify speaking rate in practice, we will distinguish and examine both pause rate (the amount of silence, not including stop closures, per second of utterance time) and articulation rate (syllables or phones per second of actual speech time, excluding silences and filled pauses). In fluent speech, it is often said that overall changes in speaking rate are largely due to changes in the pause rate, with articulation rate being relatively stable [Miller et al 1984]. In one study of conversations, the mean articulation rate ranged from 4.4 syllables/s (slowest speaker) to 5.9 syllables/s (fastest speaker); i.e., the average syllable lasted about 200 ± 25 ms. Intra-speaker variation cannot be ignored: standard deviations ranged from 0.5–1.5 syllables/s (i.e., about 20%). Dividing up speech into "runs" (periods of speech bounded by pauses), another study found an average of 11.68 syllables/run [Miller et al 1984]. For most speakers, the difference in average syllable duration between their fastest and slowest runs was quite large, exceeding 300 ms. Thus speaking rate can be very variable in conversations.

Conversational speech employs 150–250 words/min, including pauses, which average about 650 ms each [Crystal & House 1986]. Phone durations vary considerably due to factors such as speaking style (reading vs. conversation), stress, the locations of pauses and of word and syllable boundaries, place and manner of articulation, and rhythm [Summers et al 1988, Picheny et al 1986]. Most previous duration studies examine read speech and/or local (within-word) phonetic effects [Klatt 1976, van Santen 1993, Gopal 1990]. For example, in a series of papers, Crystal and House [1990] examined many local effects found in readings of two texts by six speakers. Summarizing some of their main observations here, it was found that stressed consonants (i.e., those preceding vowels in stressed syllables) averaged about 70

0-7803-2431-5/95 $4.00 © 1995 IEEE

ms, whereas unstressed consonants averaged 50 ms. Vowels generally were significantly longer and lengthened further immediately preceding a pause: non-prepausal (55 ms unstressed; 130 ms stressed) and prepausally (80 ms unstressed; 270 ms stressed). They did not find strong evidence of the phenomenon that phoneme durations reduce as the number of syllables in a word increases (which has been widely reported for words in simple citation-form utterances). This is just one of several examples of discrepancies between intonational results in artificial, heavily-controlled circumstances and those in more natural speech. When dealing with spontaneous speech, durational results found in very controlled contexts are often difficult to apply directly.

The difference in style between read and conversational speech can have significant durational effects; e.g., typically half of conversation time consists of pauses, compared to only 20% in read speech. Three durational phenomena common in read speech do not seem to occur in conversation [Umeda 1977]: (1) phrase-final lengthening, (2) polysyllabic shortening, and (3) consonantal effects on preceding vowels. (1) The final syllable of major *phrases* (word sequences grouped together by syntax) in English tends to be up to 200 ms longer than syllables in other positions. (2) Average syllable duration tends to decrease with more syllables in a word. (3) Vowels are longer before voiced consonants than before unvoiced ones [Klatt 1976, Luce et al 1985]. Both voiced and unvoiced speech exhibit a general *prepausal lengthening* of the last few syllables just prior to a pause, in which most phones (but not stops) lengthen compared with nonpausal contexts. The final prepausal syllable typically doubles in duration, while earlier syllables have lesser lengthening. This effect has been attributed to a slowing down of the speech in anticipation of a pause, aiding perceptual cues to syntactic boundaries, and/or additional time needed to accomplish large F0 movements that often accompany prepausal speech. In read speech, similar lengthening often occurs at syntactic boundaries without pauses.

The durations of phones are heavily influenced by stress and speaking rate. As noted above, stressed syllables are longer than unstressed ones. Stress primarily affects vowel duration, whereas syllable-final consonants have little stress variation; durations typically differ 10–20% between stressed and unstressed syllables. Stressed syllables are usually found in words considered important by the speaker for proper communication of a message. *Function words* (e.g., prepositions, articles, conjunctions, pronouns) are rarely stressed, while most *content words* (nouns, verbs, adverbs, adjectives) are stressed. Words unexpected by the listener or new to the conversation (*new information*) typically are stressed. Infrequently used words have longer durations than common words [Umeda 1977].

When a person speaks more slowly than normal, pauses account for about 80% of the durational increase (55% are new pauses and 25% are extensions of pauses at the normal rate) [Crystal & House 1986]. In read speech, unstressed syllables have been found to shorten more than stressed ones, vowels more than consonants, and unvoiced stops more than voiced ones [Miller 1981]. Many of the durational contrasts that may help cue phonemics at slow and normal rates tend to neutralize at fast rates [Port 1981].

It has been suggested that each phone has a minimal "incompressible" duration related to the speed of its articulators [Klatt 1976]. When subject to several shortening phenomena (e.g., a cluster consonant in an unstressed, polysyllabic word at a fast speaking rate), the net effect of all shortenings is less than the sum of all the individual effects. However, durations in such cases tend toward an asymptote larger than can be explained on articulatory grounds [Port 1981]. In modeling duration, there has been considerable disagreement as to whether rules for lengthening or shortening should be expressed absolutely or in percent and whether the rules should combine by addition or multiplication [van Santen 1993, Campbell & Isard 1991].

Rhythm is often noted as an important factor in interpreting durational phenomena. English is often described as a 'stress-timed' language in which stressed syllables tend to be regularly spaced in time (as opposed to languages such as French or Spanish, where syllables tend more toward equal emphasis and duration per syllable). Such isochrony is not so easily found in actual measurements, but the perceptual impression persists nonetheless, as highlighted by the tendency in English for unstressed syllables to shorten more than stressed syllables as speaking rate increases. Researchers often talk about groups of syllables forming a prosodic pattern containing one stressed syllable. Such 'stress groups' range from 1 to 5 syllables and average around 1/2 second in one study [Crystal & House 1990].

Given the large number of studies of read speech and local durational effects, we do not attempt to replicate such work; instead we concentrate on global effects in spontaneous speech. Acoustical analyses of global timing patterns in disfluent speech with a view toward application in speech synthesis or recognition are rare. (Indeed, all practical analysis of disfluent speech is very recent and found in [Shriberg et al 1992, Bear et al 1992, Nakatani & Hirschberg 1994] or our own work [O'Shaughnessy, 1992, 1994].) Other work on restarts has dwelled almost exclusively on the length of the repeated-word sequences (and occasionally on the pause duration).

3. SPEECH DATABASE and ANALYSIS METHODS

In the context of our investigation into voice dialog access to databases, we are examining an application involving a simulated travel agent. A naive user (the speaker) is given the task of arranging a trip involving air travel via commerical airlines, by verbally interacting with a "computer travel agent." Thus, the user formulates verbal questions and commands in a spontaneous fashion, as if in conversation with a travel agent. A large database of spontaneous speech (part of the ATIS - Air Travel Information System - database furnished by DARPA-NIST, and available through the Linguistics Data Consortium) was analyzed in terms of durational measurements, as well as textual analysis. For analysis purposes, hardcopy displays were made of all utterances containing restarts (as determined by listening and transcribing each utterance), in sections of 3-5 seconds at a time. Each display contained a waveform (amplitude vs. time) and a narrowband spectrogram (showing 0-2 kHz). Time resolution in these displays ranged from 44 to 78 mm/s; the frequency axis showed 39 mm/kHz. These displays were manually segmented into words and syllables.

4. ACOUSTICAL ANALYSIS RESULTS

Our results show a tendency for individual speakers to adopt a specific articulation rate in fluent speech (i.e., utterances without hesitations or false starts). Consistent durations are found in such speech, especially for brief monosyllabic function words. Stressed content words show greater variability, especially in prepausal situations.

If we restrict the timing analysis to fluent speech with no significant hesitations of any type (i.e., no intrasentential pauses, nor any major unusual elongation of words), speakers tend to retain a fixed speaking rate. A large amount of the variability among word durations can be accounted for by three factors: the number of phonemes in each word, whether it is a function word or a content word, and whether the word forms part of a common sequence of words. A good first approximation to any given word's duration is simply nX msec, where n is the number of phonemes in the word and X is an average phoneme duration (which depends on the speaker and on the speaking rate). For example, for one speaker we examined in detail, under his normal rate, $X = 80$ msec. This figure is very typical of many speakers. The phoneme duration X is reduced by about 12% for most function words (e.g., $X = 70$ msec) (although we found that the word 'the' averaged only 100 msec); X is increased by about 25% for content words containing 1–2 syllables (e.g., $X = 100$ msec). As seen in Table 1, these simple rules account for most of the word durations in fluent speech. (Again, we are not examining intraword effects, and so are not claiming that individual phonemes are 80 msec, but simply that the word durations can be predicted in terms of the number of phonemes in the word; there obviously is much variability in the durations of individual phonemes.) Table 1 shows statistics for several words that occurred often in our database (in the table, 'deviation' refers to the maximum deviation (either positive or negative) from the average duration). The results for other speakers were quite similar, with small differences in the values for X and for the adjustments for function words (e.g., –12%) and for short content words (e.g., +25%).

Smaller durational effects for each word can be attributed to the semantic importance of the word. For example, there was a tendency for words at the start of a syntactic unit to be slightly shorter than those at the end of the unit (ignoring cases of prepausal lengthening, where very substantial lengthening occurs on the last word before a pause). In addition, in each utterance, there is often one key word which conveys the most important semantic information of the sentence; this word typically lengthens by up to 200 msec beyond what would be predicted by the normal durational rules. Conversely, groups of words which form common expressions (e.g., in our database, 'show me all the ...' or 'give me the ...') tended to have slightly shortened durations. For example, many utterances in this database started with 'What are the...'; a 3-phoneme word such as 'what' would normally have about 210 ms duration (indeed this range was found in the database in other contexts), but in this redundant context, it averaged 133 msec.

Similar shortening occurs for repeated content words: the first time a content word is introduced into a discourse, it is stressed (often with extra duration, as noted above); repeated occurrences of such a word in the immediately ensuing discourse have shortened durations. Lastly, there were several cases of lengthened function words, where the speaker slowed down his articulation during the function word, apparently while searching for the proper content word to follow. The elongation was not so much to label it as a hesitation, but clearly when the speaker is not totally sure of what he is saying from the start of the utterance, there will be places where he slows down slightly to think, and these tend to occur on function words.

Word	Average	Phonemes	Deviation
the	100	2	50
all	175	2	75
on	125	2	25
in	125	2	75
me	150	2	50
from	250	4	100
to	125	2	25
what	133	3	40
are	80	1	20
information	625	9	75
San Francisco	850	12	50
Baltimore	515	8	35
Dallas	425	5	75
Atlanta	575	7	75
Boston	550	6	90
nonstop	690	7	60
flight	400	4	50
flights	470	5	100

Table 1 Word duration statistics for one speaker (in msec).

Prepausal lengthening is highly variable, but mostly limited to the word immediately before a pause. In a minority of the cases, the word prior to the pause is not lengthened at all (i.e., the speaker effectively abruptly stops speaking, without any anticipatory slowing down in the prior speech). Most of the time, the final word is lengthened significantly, by 100-300 msec typically (even for function words).

The greatest variability in speech timing occurs in disfluent speech, where hesitations and false starts cause extremes in slow and fast speech. Specifically, upon reaching a hesitation point, the typical speaker either: 1) abruptly stops (switching quickly from a normal articulation rate to a pause), 2) abruptly slows down for 1–2 syllables (often followed by a pause), or 3) enters a mode of much slower speech for a few words (often containing pauses). Speech after the pause fell into two categories: 1) normal rate speech (i.e., a return to normalcy), 2) accelerated speech for the repeated words.

Just looking at silent pauses in otherwise fluent speech (i.e., ignoring filled pauses and pauses occurring in restarts [O'Shaughnessy 1992, 1994]), there was still a wide variability in pause durations. Ones occurring within small syntactic units (such 'ungrammatical pauses' are never heard in fluent speech) tended to be short, averaging about 150 ± 100 msec; about 70% of such pauses were in this range, with the other 30% equally spread over the range 300-1000 msec. Speakers try to minimize the duration of such unexpected pauses, and succeed in most cases in restarting the verbal

communication with a very brief delay (equivalent to only about two phonemes' duration).

The grammatical pauses (i.e., ones at major syntactic boundaries, which often occur in lengthy fluent sentences), on the other hand, did not concentrate in one durational range. It appears that each speaker has certain preferred pause ranges. Our main speaker preferred three relatively constrained ranges (125 ± 25, 375 ± 25, and 850 ± 150 msec), where 58% of his grammatical pauses occurred. One could speculate that such durations reflect typical thought delays in the cognitive processes of speech production: the brief one being quite similar to most hesitation pauses, the longer one being used when a bit more thought is required, and the longest one when major sentence organization is required mid-utterance.

Speaking rate was sometimes adjusted at the time of a pause, in the sense that prior to the pause the speaker went more slowly than normal (presumably actively thinking ahead while talking) and then returned to his normal speaking rate after the pause (which was used to complete the planning thoughts). Typically, durations in all words prior to such a pause were about 60% longer than afterward. Effectively, this is prepausal lengthening, except that normally such lengthening only affects the word just prior to the pause.

With false starts, when a word was simply repeated (as is) in a restart, it had virtually the same duration in both its instances in most cases. When a word was changed (i.e., a substitution or insertion) in the restart, on the other hand, its second instance was virtually always more stressed (i.e., longer duration). The substituted/inserted word received a large stress (relatively long duration) in examples where the new word added significant semantic information, but did not in examples where the new word was redundant in terms of the prior context (e.g., if the new word was a synonym of an immediately previous word). As for the repeated words (after the pause) prior to the inserted word, function words showed little or no shortening; on the other hand, content words here exhibited significant shortening (the shortening here was about 50% for short words less than 300 ms, and about 100-200 ms for longer words). Such prosodic change only applied to non-prepausal words, because words immediately prior to a pause were often subject to significant prepausal lengthening.

5. CONCLUSION

We have described aspects of timing in spontaneous speech, specifically for global timing patterns. We detailed prosodic phenomena in fluent and disfluent utterances in a database of spontaneous, continuous speech, and gave intuitive explanations for them, based on a theory of using prosodics to cue semantic information to a listener. Based on the acoustic data, the models described can explain much of these timing patterns.

6. REFERENCES

Bear, J.; Dowling, J.; Shriberg, E.: "Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog." Proc. Assoc. Computational Lingusitics, 56–63, June 1992.

Campbell, W. & S. Isard: "Segment durations in a syllable frame," J. Phonetics, vol. 19, 37–47, 1991.

Chafe, W.: "Some reasons for hesitating," *Temporal variations in speech* (H. Dechert & M. Raupach, eds.), 169–180, 1980.

Crystal, T. & A. House: "Characterization and modeling of speech-segment durations," Proc. IEEE Int. Conf. ASSP, 2791–2794, 1986.

Crystal, T.; House, A.: "Articulation rate and the duration of syllables and stress groups in connected speech," J. Acoust. Soc. Am., vol. 88, 101–112, 1990.

Gopal, H.: "Effects of speaking rate on the behavior of tense and lax vowel durations," J. Phonetics, vol. 18, 497–518, 1990.

Klatt, D.: "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am., vol. 59, 1208–1221, 1976.

Luce, P. & J. Charles-Luce: "Contextual effects on vowel duration, closure duration and the consonant/vowel ratio in speech production," J. Acoust. Soc. Am. **78**, 1949–1957, 1985.

Miller, J.: "Effects of speaking rate on segmental distinctions," in *Perspectives on the study of speech* (P. Eimas & J. Miller, eds.), 39–74, 1981.

Miller, J., F. Grosjean, & C. Lomanto: "Articulation rate and its variability in spontaneous speech." Phonetica, vol. 41, 215–225, 1984.

Nakatani, C. & Hirschberg, J.: "A Corpus-based study of repair cues in spontaneous speech." J. Acoust. Soc. Am., vol. 95, pp. 1603–1616, 1994.

O'Shaughnessy, D.: "Correcting complex false starts in spontaneous speech." ICASSP-94, 349–352, 1994.

O'Shaughnessy, D.: "Automatic recognition of hesitations in spontaneous speech." ICASSP-92, 593–596, 1992.

Picheny, M., N. Durlach & L. Braida: "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. Speech Hear. Res. **29**, 434–446, 1986.

Port, R.: "Linguistic timing factors in combination," J. Acoust. Soc. Am. **69**, 262–274, 1981.

Summers, W., D. Pisoni, R. Bernacki, R. Pedlow & M. Stokes: "Effects of noise on speech production: Acoustic and perceptual analyses," J. Acoust. Soc. Am. **84**, 917–928, 1988.

Shriberg, E.; Bear, J.; Dowling, J.: "Automatic Detection and Correction of Repairs in Human-Computer Dialog." DARPA Speech and Natural Language Workshop, Feb. 1992.

Snidecor, J.: "A comparative study of the pitch and duration characteristics of impromptu speaking and oral reading," Speech Monographs, vol. 10, 52–56, 1943.

Umeda, N.: "Consonant duration in American English," J. Acoust. Soc. Am. **61**, 846–858, 1977.

van Santen, J.: "Timing in text-to-speech systems," Proc. EUROSPEECH-93, Berlin, 1397–1404, 1993.