Timing in conversations: from speech synthesis to robot Interaction

Wzorce Czasowe w konwersacjach: od syntezy mowy do interakcji z robotem

Nick Campbell

Speech Communication Laboratory
Centre for Language and Communication Studies
Trinity College Dublin, Ireland
nick@tcd.ie

ABSTRACT

Speech synthesis is now a mature technology, and many issues that were of concern to earlier generations have now become well understood, but for use in conversational interaction, the technology is still inadequate. Similarly, whereas phonemic timing and sentence-internal durational dependencies are now also well understood, the timing of utterance sequences in an interactive conversation is still an area rich for future research. This paper explores higher-level speech timing using human-robot interactions and suggests that the framework elaborated by Jassem may hold true for even higher levels of speech timing structure.

STRESZCZENIE

Synteza mowy jest obecnie dojrzałą technologią i dziś już dobrze rozumiemy wiele spośród kwestii, które były problematyczne dla wcześniejszych pokoleń, ale do zastosowań w interakcji konwersacyjnej technologia pozostaje wciąż nieadekwatna. Podobnie, podczas gdy dobrze rozumiemy realizację czasową głosek i zależności iloczasowe na poziomie zdania, realizacja czasowa sekwencji wypowiedzi w interaktywnej konwersacji nadal stanowi bogate pole dla przyszłych badań. Niniejsza praca bada realizację czasową mowy na wyższym poziomie poprzez wykorzystanie interakcji między człowiekiem i robotem oraz pokazuje, że schemat wypracowany przez Jassema sprawdza się nawet dla wyższych poziomów struktury czasowej mowy.

1. Early work on timing for computer speech synthesis

The work of Wiktor Jassem had a great influence on my own early research into explaining and predicting the timing of individual speech sounds for the computer synthesis of English and Japanese. An extension of his ideas regarding speech rhythm can now perhaps be extended to higher-level units in conversational interaction, as I shall propose later in this chapter.

My thinking in the early eighties, at the time I was working on predicting segmental durations for speech synthesis, was of course also heavily influenced by the seminal work of Denis Klatt [1], and I started by incorporating his duration rules into a prediction

engine for English speech rhythm. They worked well at the individual sentence level but did not generalise well to speech in a broader context since they produced only one fixed duration for each sound in a given text sequence, regardless of any difference or variation in the intended interpretation of that text.

Halliday had shown in his description of English phonology [2] that we need to recognise four rhythmic units: in descending order, they are the tone group, the foot, the syllable and the phoneme. Abercrombie [3] defined the foot as a 'unit which starts with a stress and includes everything up to, but not including, the next stress, and within which, rhythmically, the syllable functions'. Jassem and Gibbon [4] might prefer the use of the word 'accent' here, in place of 'stress'. The foot is independent of word boundaries. Abercrombie [3:p28] defines syllable quantity as neither directly dependent on vowel quality nor on stress, but rather as a proportion of the total length of the foot within which the syllable occurs, and therefore relative to the quantity of any other syllable in that foot. My own prediction model was based on this concept of elasticity of segments accommodating to constraints within a higher-level timing framework.

Pauses had syllabic status in this timing framework and were known to function as stressed syllables with respect to foot structure. Abercrombie cited Steele 'regarding a line of verse as perfectly regular when it was short of the number of *stressed syllables* (Abercrombie's italics) which in theory it ought to contain, but had pauses for a compensating number of stresses to fall on', and he drew a parallel [3, 5:p20] with the *silent stress pulse* (Abercrombie's italics) that he claimed to exist in spoken prose. In Halliday's summary of Abercrombie's model a foot is taken to start with a silent stress if it lacks an initial stressed syllable and follows a pause or has initial position in the tone group.

For Jassem, "English speech consists of two kinds of rhythm units: (a) Narrow Rhythm Units (NRU) and (b) Anacruses (ANA) [6]. For a given tempo, the length of a narrow rhythm unit depends on the number of syllables. This length is a constant for a monosyllabic rhythm unit and a given tempo, ... As the number of syllables in a narrow rhythm unit increases, the length of the narrow rhythm unit (NRU) also increases, but not proportionately."

D. R. Hill, in his 'Conceptuary for speech and hearing in the context of machines and experimentation' [7], cf. also [8], summarises Jassem's view of speech rhythm as follows: "In a stress-timed language such as English, the speech may be divided into rhythmic units related to the occurrence of the stresses, or beats. Halliday, following David Abercrombie, puts the rhythmic unit boundaries just before each syllable bearing primary stress and calls the unit a 'foot' [2]. Jassem's scheme is similar, but excludes proclitic syllables from consideration, likening them to the anacruses in music. Such frameworks can provide a basis for workable models of English rhythm, *but require extensive analysis of real speech segments*. One such study by Hill together with Jassem and Witten slightly favoured the Jassem formulation, but a subsequent model of rhythm for use in speech-synthesis-by-rules used the Halliday formulation because it is "a little simpler" in not requiring grammatical information [9; 8].

Being with IBM at that time, as a guest researcher in their UK Scientific Centre, I had access to a large corpus of recordings of spoken English, and spent considerable time measuring individual syllable durations in a variety of speech contexts. My goal was to predict the segmental durations within each syllable. I was able to improve on

the predictions of the Klatt model by incorporating Jassem's notion of the foot, with its anacrustic variant, as a level of timing control in my statistical model. Jassem's framework for speech rhythm introduced a layer of control and some distinctions that Klatt had not taken onto account. His distinction of unstressed initial syllables (anacrusis) also provided a better fit to the rhythmic structure of the corpus data [10].

1.1. Multi-level timing control in English speech

On the basis of this framework, I developed a statistical model of speech timing that employed neural networks (which were just becoming popular at the time) to accommodate segmental durations by a process of optimisation between a higher 'cognitive' level of timing control, and a lower 'mechanical' level of timing production.

Jassem claimed that "The rhythm of English speech is a phonetic phenomenon and is determined on purely phonetic principles with no recourse to any other level of analysis-synthesis, such as grammar or semantics" (point B.6 in [6]), but I found that being able to incorporate a higher cognitive level into the model, which then allowed for pragmatic differences in expression, provided a better fit to the corpus material; cf. also Hill's 'quibble' with Jassem [8]. In the default case, however, the structure of the text and the sequence of phones in a given prosodic framework would have been sufficient to predict segmental durations with sufficient accuracy to produce natural-sounding synthesised speech.

The cognitive aspect allowed for different speaking rates, and different pragmatic interpretations of an otherwise identical word sequence or sentence, allowing for the differences in phrase boundary placement that result from different semantic interpretations. The 'mechanical' aspect, on the other hand, was directly related to the physics of speech production and the generation of the individual phones in the speech sequence. The resolution of these different influences on segmental timing was carried out using a concept of 'stretch' whereby inherently longer phone sequences would expand or shrink to fit the syllable framework according to a phone-type-specific factor of elasticity.

1.2. Timing control in read speech

Early timing research was carried out using recordings of read speech, with most studies, extending even up to the time of Klatt, basing their measurements on spectrograms of recorded speech. Each utterance was produced in isolation and the sentences were also typically very short because of the limits of the recording medium. The recordings were usually made in a laboratory, where the main task of the reader was simply to render the text into clear and intelligible speech. There is very little pragmatic variation in such recordings as the relationship of the reader to any future listeners is undefined. The role of duration in such read-speech material is primarily to distinguish phonetic variants and to mark phrasal boundaries so that the linguistic meaning of the underlying text is made clear. Interactive speech, on the other hand, is far more variable.

Figure 1 (reproduced from [11]) clearly shows this difference in the variability of segmental durations observed in three reading styles (isolated words, sentences, and continuous or contextualised sentences) and in spontaneous speech where interaction with the interlocutor has a greater effect on the situated or context-dependent utterances, as seen in the vertical expansion (more variability) versus horizontal compression (little variability in the means).

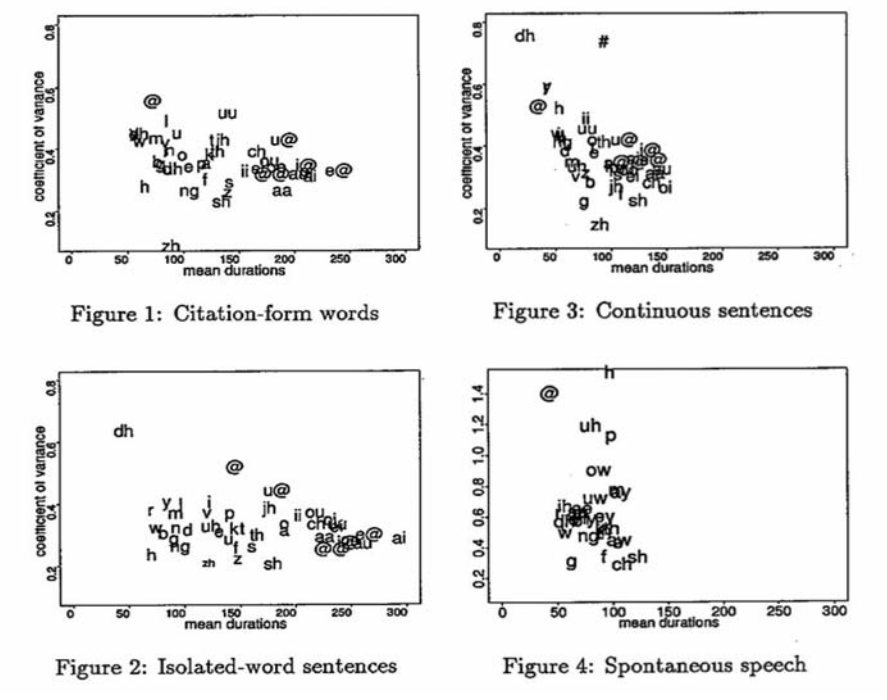

Figure 1: Four figures from [11] showing the differences in segmental variation between continuous and isolated readings.

2. Timing models for interactive speech synthesis

Now that computer speech synthesis has become a mature technology, being used in a variety of everyday applications, the prediction of segmental timings for broadcast-type speech (as used in announcements or reading and in contrast to interactive or conversational speech) can be considered sufficient for purpose. Researchers can now begin to model the next higher level of timing control; that between utterances in a conversation.

Future speech synthesis must be interactive, not just reading-out text, but talking to a person as part of an interactive dialogue system. The synthesiser will need to be sensitive to the interlocutor's reactions and should time its utterances accordingly. Research for such systems must be based on corpora of actual speech, and the focus must shift from individual sentences to entire utterances. The granularity has changed, but the relative structure can be considered similar, and we will see that Jassem's insights into the nature of the timing framework hold true for this higher level too.

Our present corpora include audio and visual face-to-face as well as telephone conversations. In this paper we focus on the latter since for our face-to-face daily-life interaction data, we only have the right to study the speech of the microphone-wearer as there are no contractual agreements with the people they spoke to while going about their daily routines.

Figure 2: Differences in conversational style in telephone conversations, as illustrated by plots of speech activity, showing a balanced turn-taking mode (top part) contrasted with a more fragmented interaction mode (lower part).

The telephone corpus [12] was recorded over a period of several months by pairs of people balanced for age, sex, and familiarity. Analysis of speech activity revealed at least three modes of interaction, depending on who spoke when, for how long, and with what degree of overlap.

Figure 2 shows a plot of speech activity patterns from the telephone recordings. Each bar represents the vocal activity of one speaker. Each line shows a minute of conversation. The top part of the figure is illustrative of a balanced conversation where each speaker dominates in turn, and the lower part shows a more social mode of interaction where both speakers employ a fragmented style, and much laughter is produced as seen by the shorter bursts of speech or vocal activity. The types of timing patterns shown here are of great interest. Our goal is to model social aspects of speech interaction as illustrated by these different modes of dialogue activity, but the present work focusses on predicting the length of the utterances (bars) and the timing of their onsets in a more predictable type of interaction. We therefore recorded a corpus of people talking with machines, where the social and interpersonal aspects are minimised and a more formal mode of interaction can be expected.

3. Current work on timing for interactive speech conversation

From April to June in 2011, visitors to the Science Gallery (a high-tech exhibition space) in Dublin came across an exhibit featuring a small LEGO robot ("Herme" [13]) that started a conversation when they approached. The nature of the conversation (shown below) was such that the majority of people cooperated and gave formal consent to the audio-visual recordings of their conversation being used for further research.

3.1. Meta-Feet, functioning at the discourse level

The conversation that we designed for Herme the robot evolved as our experience grew throughout the exhibition, but stabilised rapidly as we discovered a routine or sequence of interactions that worked well with most people. Perhaps the dialogue engaged interlocutors so well because it involved 'meta-feet', or groups of utterances with a heavy onset and typically a pair of weaker follow-ons. In the script used by Herme the robot to talk with exhibition visitors each meta-foot is indicated by a dash:

```
– hello? Hi ...  hello, hi.

– my name is Herme – h e r m e – 'herme',
– whats your name?

– how old are you?
  really, I'm nearly seven weeks old

– do you have an i-d number
  I need an i-d number to talk to you
  i d numbers are on your right . . . . thank you

– are you from dublin?
  really
– I'm from the Speech Communication Lab here in TCD
  tell me about you...
  really? oh

– tell me something else
– why you here today?
  oh really
  why

– do you like the exhibition
  really
  why?

– i like your hair!

– do you know any good jokes?
  tell me a funny joke
  ha ha haha ha

– tell me a knock knock joke
– who's there
  who?
  who
  ha ha haha ha
```

```
– I know a joke
– what's yellow and goes through walls ?
  a ghost banana – ha ha hehe he
  ho hoho ho ho

– thanks for your help goodbye
  see you later – goodbye
```

It was important for the robot to maintain control of the turns in the conversation because speech recognition is not yet strong enough to provide a reliable estimate of what each visitor said in the noisy environment.

Each turn consisted of a comment or question that was followed-up by nonverbal backchannel feedback or laughter. The initiation served to draw the visitor into the conversation (Figure 3). Two modes were used in collecting the corpus; (i) fully-automatic dialogue, speaking according to the script, with timings triggered by voice and image processing using routines programmed in MAX/MSP, or (ii) a wizard-mediated version triggered by use of Skype using the iSight camera and microphone that can be seen beneath the display in Figure 3. The wizard was a researcher situated in a separate building who could both see and hear the visitor's response and who was able to signal to the robot when to speak. In both cases, the dialogue sequence was identical and the only difference was in the timing of the initiation of each utterance. It will perhaps come as no surprise that the conversations mediated by the wizard were more successful, resulting in more consent forms signed.

Figure 3, from [14] shows the clustering of gap timings resulting from the use of the wizard as a trigger. Units are not relevant here, and the focus of the figure is on the tri-modal clustering of the peaks into short (for the backchannel comments), medium (for the inter-group pauses), and long (when the visitor is providing personal information).

Figure 3: A visitor to the Science Gallery in Dublin taking part in a conversation with the robot. She can see that the robot is looking at her and responds naturally to its questions as it leads her through the conversation.

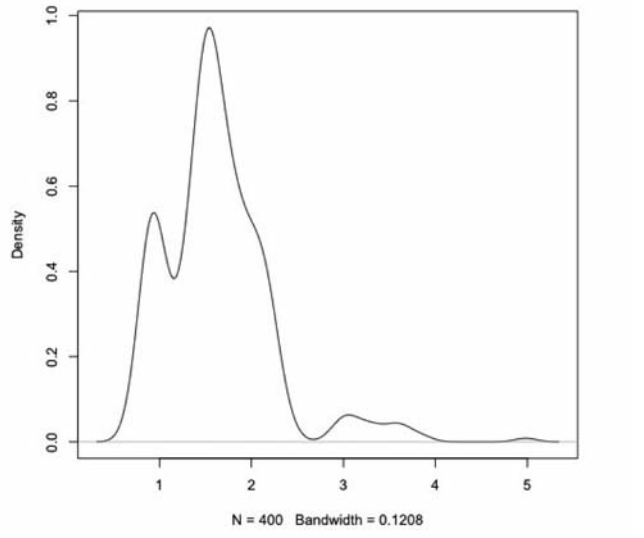

Figure 4: Gap timings in utterance-triggering when per-formed by the wizard.

Notice how the distribution of the three peaks corresponds so well with observed distributions of syllable duration at the next level down in the hierarchy of speech unit timings.

4. Conclusion

At TCD, we recently implemented a conversational robot which, over a period of three months, engaged passers-by in small-talk at a Science Gallery exhibition. Typical conversations lasted from three to five minutes, during which interactants were asked (by the robot) to sign a consent form allowing the video and audio recordings of their conversation to be used in a corpus of naturally-occurring human-machine speech. Almost five hundred visitors signed, indicating a successful completion of the interaction, but more than seven hundred interlocutors walked away before completing the dialogue. Analysis of the successful versus the unsuccessful conversations shows utterance timing to be the main factor that predicts failure of an interaction. We describe this work and claim that conversational utterances display a foot-like structure important for social cohesion. The robot fails to convince the human partner that it is actively participating in the conversation if it mis-times the start of an onset utterance.

The study of timing in speech continues to be of great importance both to further our knowledge of how human interaction works, and to provide technology for unobtrusive, easy to use, speech-based human-machine interfaces for future dialogue systems.

This paper has presented an overview of some early and some very recent work on the prediction and analysis of timing in speech, and has shown that the seminal work of Jassem is still of relevance and can probably be extended to even higher levels of timing control.

We acknowledge that our work on timing in conversational speech is still very premature but wish to take this opportunity to recognise the influence of Wiktor Jassem and to show with very recent corpus material that his insights into and understanding of the processes of speech timing may have deeper repercussions than we once thought.

Acknowledgement

This work was undertaken as part of the FASTNET project – *Focus on Action in Social Talk: Network Enabling Technology*, funded by Science Foundation Ireland (SFI) 09/IN.1/I2631. We thank the Digital Hub and the Science Gallery at TCD for the opportunity to display the robot in the Human+ exhibition.

References

[1] Klatt, D. H. 1979. Synthesis by rule of segmental durations in English sentences: In B. Lindblom and S. Öhman, eds. *Frontiers of Speech Communication Research*. New York: Academic Press. 287–300.

[2] Halliday, M.A.K. 1967. *Intonation and Grammar in British English*. The Hague: Mouton.

[3] Abercrombie, D. 1964. Syllable quantity and enclitics in English. In D. Abercrombie et al., eds. *In Honour of Daniel Jones*. London: Longmans. 216–222.

[4] Jassem, W. and Gibbon, D. 1980. Re-defining English stress. *Journal of the International Phonetic Association* 10, 2–16.

[5] Abercrombie, D. 1965. *Studies in Phonetics and Linguistics*. London: Oxford University Press.

[6] Jassem, W. 1952. *Intonation of Conversational English (Educated Southern British). Prace Wrocławskiego Towarzystwa Naukowego (Travaux de la Société des Sciences et des Lettres de Wrocław)*. Seria A. Nr. 45. Wrocław: Nakładem Wrocławskiego Towarzystwa Naukowego.

[7] Hill, D. R. 1990. A Conceptuary for speech & hearing in the context of machines and experimentation. *Technical Research Report*: http://pages.cpsc.ucalgary.ca/~hill/papers/conc/index.htm.

[8] Hill, D. R. 2012. From research to application: creating and applying models of British RP English rhythm and intonation. In D. Gibbon, D. Hirst and N. Campbell, eds, *Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem*. Poznań: Polskie towarzystwo Fonetychne/Polish Phonetics Association. (This volume.).

[9] Jassem, W., D. R. Hill and I. H. Witten. 1978. *A statistical approach to the problem of isochrony in British English*. Research Report No 78/27/6, University of Calgary, January 1978.

[10] Campbell, W. N. *Multilevel Timing in Speech*. 1990. Unpublished PhD thesis, University of Sussex, U. K.

[11] Campbell, N. 1995. From read speech to real speech. Invited paper: Symposium on the phonetics and phonology of speaking styles. *Proc, XIIIth International Congress of Phonetic Sciences*, Stockholm.

[12] Campbell, N. 2004. Databases of Expressive Speech. *Journal of Chinese Language and Computing*, 14 (4), 295–304, 2004.

[13] Vaughan, B., J. G. Han, E. Gilmartin and N. Campbell. Designing and Implementing a Platform for Collecting Multi-Modal Data of Human-Robot Interaction. *Acta Polytechnica Hungarica*, 9 (1), 7–17.

[14] Gilmartin, E., C. De Looze, and N. Campbell. Priming, Timing, and the Phatic Component in Machine-Mediated Dialogue. In *Proceedings of the LISTA workshop*, Edinburgh, May 2012.