PERCEPTION OF SPEECH RATE AND NATURALNESS IN SYNTHETIC SLOW SPEECH

Cyril Auran & Caroline Bouzon

Laboratoire Savoirs Textes Langage, CNRS UMR 8163,
Université Lille 3 - Charles-de-Gaulle, Lille, France
cyril.auran@univ-lille3.fr; caroline.bouzon@univ-lille3.fr

ABSTRACT

This paper details two perception experiments based on synthetic British English obtained with CART models predicting phone durations in slow speech from normal speed speech. Speech rate and naturalness were assessed by 6 English natives. Synthetic slow speech was rated as both slower and less natural than natural slow speech; however, the insertion of the pauses produced in natural slow speech into the synthetic recordings suggests a better approximation of natural speech in terms of both naturalness and perceived speech rate.

Keywords: British English, speech rate, perception, naturalness, speech synthesis

1. INTRODUCTION

The research described here fits in a more global project concerning the modeling of orthogonal prosodic dimensions in discourse ([3, 13]). These orthogonal prosodic dimensions have been shown to play an important part in the marking of the topical structure of discourse, for instance with increased pitch level and range signaling the beginning of a new topic, and lower pitch level and range ("final lowering") at the end of topics ([9, 21]). More specifically, our current research focuses on speech rate in British English. With relation to discourse structure, slower speech rate is involved in the marking of topic beginnings ([12, 18]) and topic ends together with final lengthening ([18, 21]).

This paper builds upon the Classification And Regression Tree (CART) predictive model presented in [4], which can be used to generate artificially slowed down utterances from "normal speech rate" input. Our study thus constitutes a follow-up to the quantitative (objective) evaluation given in [4] and more specifically assesses the (subjective) perception of slow speech synthesized with this model in terms of speech rate and naturalness.

Section 2 provides an overview of the synthesis procedure, with particular focus on the CART models and their implementations. Sections 3 and 4 detail the perception experiments and section 5 provides a temporary conclusion and perspectives.

2. SYNTHESIS

2.1. CART models

Classical regression techniques, such as linear models, are not easily interpretable with the kind of complex patterns reported in [4], with interactions between such parameters as Speaker, Phase (normal *vs.* slow speech), Stress or Position within the Inter-Silence Segment (ISS). Classification And Regression Trees (CARTs; [8]), may thus be preferred, with the advantage of selecting the most significant parameters, providing "honest" estimates of their performance, allowing both categorical and continuous features to be considered and allowing straightforward interpretation of the results (see [17] for an illustration in segmental duration modeling).

2.1.1. Generation

In [4], the CARTs were generated within the R environment using the rpart package ([19]). Tree over-fitting was controlled through cost-complexity pruning and cross-validation over the entire data set (thresholds reminded below), a method proposed in [8] which minimizes the variance of the prediction error as a function of tree length. The training and predictions of the CARTs were carried out on the same data set (20 utterances by three different speakers).

Two groups of CARTs predicting phone durations were produced for each speaker, with further distinction between Abercrombie's [1] and Jassem's [11] rhythmic models. This allowed a comparison of the models (which differ in the definition of their rhythmic units) in terms of phone duration predictions. The initial parameter set was restricted to Stress, Rhythmic unit type

(Abercrombie's model *vs.* Jassem's model), Number of constituents and Position (from beginning and from end) in the ISS, the rhythmic unit and the syllable.

2.1.2. Quantitative evaluation of the predictions

[4] provides results for phone durations for a single speaker (Speaker F) in both normal and slow speech, with a comparison of the contributions of Abercrombie's and Jassem's models. Optimised CARTs (normal and slow speech) were generated with Abercrombie's and Jassem's models through cost-complexity pruning with thresholds of 0.021 (normal speech; Abercrombie's model), 0.024 (normal speech; Jassem's model), 0.023 (slow speech; Abercrombie's model), 0.023 (slow speech; Jassem's model).

As can be seen in table 1, both models provide a mean absolute error of about 13 ms in normal speech and 16 ms in slow speech (analogous to the values given in [17] with a similar method). However, CARTs generated using Jassem's model display lower complexity (fewer splits).

Table 1: CART mean absolute error and split complexity (speaker F).

Speech rate	Rhythmic Model	Mean abs. error	Split complexity
Normal	Abercrombie's	12.5	5
	Jassem's	13	3
Slow	Abercrombie's	15.5	4
	Jassem's	16	2

2.2. Implementation

Speech signals corresponding to the CARTs generated for speaker F were synthesized from natural "normal speed" recordings using TD-PSOLA [14] resynthesis within Praat v 5.2.15 [5]. The minimum and maximum F0 values used for PSOLA resynthesis were computed using the quartile-based methodology provided in the Momel Praat plugin ([10]).

A total of 4 synthetic recordings were generated for each utterance: Abercrombie-based or Jassem-based synthetic slow speech, with and without the pauses from the speaker's original slow recording (see [4] for a detailed analysis of pauses).

3. PERCEPTION TEST 1: SPEECH RATE

The purpose of this first test was to evaluate the speech rate of slow synthetic recordings as perceived by natives in relation to both normal and slow natural recordings.

3.1. Method

The experiment was conducted in a quiet room in Lille 3 University on a Dell Latitude E5500 computer running Perceval Software v 3.0.5 ([2]) under MS Windows XP SP3. The stimuli were heard through Sennheiser HD 280 Pro headphones and subjects responded using the Perceval Button box.

After a short training session, subjects listened to the recordings and were instructed to rate them in terms of speech rate on a 5-point scale ranging from "very slow" (1) to "normal" (5); the absence of fast speech was explicitly mentioned. The instructions were displayed on the computer screen during the whole experiment.

The audio stimuli consisted of 120 (6 versions of 20 different utterances) PSOLA resynthesized 44.1 kHz mono wav files:

- Version A: Natural, normal speech
- Version B: Natural, slow speech
- Version B_A1: Synthetic, slow speech; Abercrombie model; source: version A
- Version B_A2: Synthetic, slow speech; Abercrombie model; sources: version A + pauses from version B
- Version B_J1: Synthetic, slow speech; Jassem model; source: version A
- Version B_J2: Synthetic, slow speech; Jassem model; sources: version A + pauses from version B

Versions A and B were resynthesized without any prosodic changes in order to minimize any perceptive bias due to PSOLA artefacts.

Each stimulus consisted of a single utterance (5 to 13 words, 8 to 25 syllables) originally read by a female British English native speaker (see [6] for a description of this sub-corpus).

The subjects were 6 native speakers of British English (5 female, 1 male; 1 left-handed), aged between 20 and 23, and did not report any hearing problems.

All the statistical analyses were carried out within R (base package and packages irr and mclust), and showed no influence of gender or lateral preference on the responses.

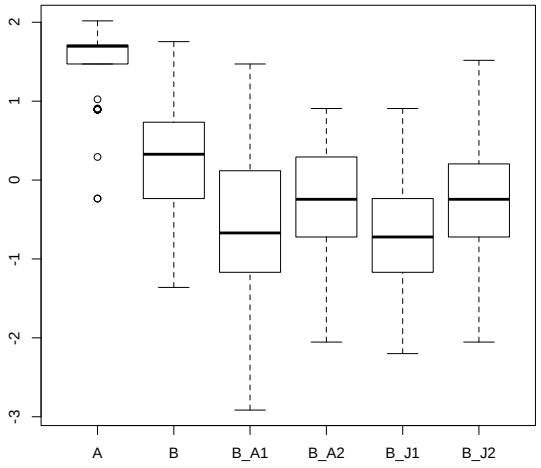
3.2. Results

The first interesting result concerns inter-rater agreement. We found significant differences between subjects in terms of responses ($F(5,714)=20.37$, $p<0.001$; $ICC(A,1)=0.56$;

ICC(C,1)=0.66), with subjects 1 and 6 on the one hand and 2, 3 and 5 on the other pooling together, and subject 4 an intermediate rater (TukeyHSD test). The specificity of subjects 1 and 6 lay in an increased use of lower values, and an overrepresented central response (level #3) for subject 4. These differences were neutralized using z-score standardization for each rater ($F(5,714)=1.895e-06$, $p=1$).

Among the factors analyzed in this study (version, order of presentation, version of the previous stimulus, reaction time and their interactions), version, unsurprisingly, was the only significant factor ($F(5,642)=189.95$, $p<0.001$). Figure 1 represents the standardized responses by version. Post hoc analyses (TukeyHSD) showed that version A (natural normal speed speech; mean z-response=1.557) was rated significantly higher than all other versions; version B (natural slow speech; z-response=0.255) was rated significantly higher than all other (synthetic) slow speech versions. The synthetic versions including version B pauses (B_A2 and B_J2) did not significantly differ from each other (TukeyHSD $p=0.98$), and display higher (though non-significant) values than the synthetic version based on version A only (B_A1 and B_J1).

Figure 1: Z-standardized responses by version.



Analyses carried out on reaction time showed significant differences between raters ($F(5,305)=11.22$, $p<0.001$), which were neutralized using z-score standardization ($F(5,642)=0.02$, $p=0.99$). Of all the factors considered, order of presentation of the stimuli was the only parameter to significantly impact reaction time ($F(1,718)=11.83$, $p<0.001$); with an adjusted r^2 of 0.015, and a coefficient of -0.004, however order of presentation only weakly diminished reaction time as the experiment unfolded.

4. PERCEPTION TEST 2: NATURALNESS

The purpose of this second test was to evaluate the naturalness of slow synthetic recordings as perceived by natives in relation to both normal and slow natural recordings.

4.1. Method

Experiments 1 and 2 were carried out sequentially. Experimental conditions, equipment and subjects were identical in both cases.

After a short training session, subjects listened to the stimuli and were instructed to rate them in terms of naturalness on a 5-point scale ranging from "very unnatural" (1) to "(close to) natural" (5). The instructions were displayed on the computer screen during the whole experiment.

The audio stimuli consisted of 100 (5 versions of 20 different utterances) PSOLA resynthesized 44.1 kHz mono wav files. These files were identical to those used in experiment 1, with the exception of version A files, which were not used in this experiment, thus providing natural (version B) and synthetic (versions B_A1, B_A2, B_J1 and B_J2) slow speech stimuli only.

As in the first experiment, version B was resynthesized without any prosodic changes in order to minimize any perceptive bias due to PSOLA artefacts.

4.2. Results

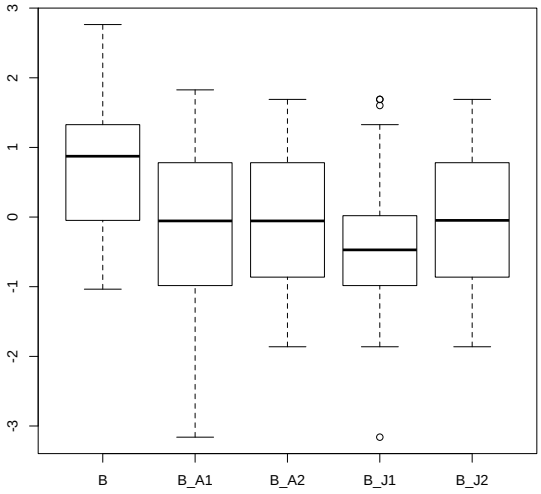
Here too, inter-rater agreement was shown to be rather moderate (ICC(A,1)=0.36; ICC(C,1)=0.46) with significant differences between raters ($F(5,594)=28.74$, $p<0.001$). TukeyHSD analyses showed that subjects pooled into 2 sub-groups: subjects 1 and 5 (with more low ratings) and subjects 2, 3, 4 and 6. Here again, these differences were neutralized using z-score standardization ($F(5,594)= 3.083e-06$, $p=1$).

As can be seen in figure 2, only stimulus version was found to significantly influence z-standardized responses ($F(4,544)= 28.63$, $p<0.001$). Version B (mean=0.78) was rated as more natural than all the synthetic versions (TukeyHSD $p=0$). Only B_J1 (mean=-0.39) differed from the other synthetic versions.

Analyses carried out on reaction time showed significant differences between raters ($F(5,305)=4.7$, $p<0.001$), with subject 6 displaying

significantly higher values (TukeyHSD $p<0.05$). These differences were neutralized using z-score standardization ($F(5,305)=0.004$, $p=1$). Stimulus version was the only factor to weakly influence z-standardized reaction time ($F(4,305)=2.48$, $p<0.05$), with reaction times for version B_J1 significantly higher (TukeyHSD $p<0.05$) than for version B.

Figure 2: Z-standardized responses by version.



5. CONCLUSION AND PERSPECTIVES

CART-modeled slow synthetic speech generated from normal speed recordings was judged significantly slower and less natural than human "slow speed" utterances. No differences were found between stimuli based on CARTs using Abercrombie's or Jassem's models, thus further confirming ([6, 7]) the interest of Jassem's approach to rhythm in English.

Our results also suggest that the inclusion of "natural" pauses in synthetic speech induces a closer approximation of natural speech in terms of speech rate and naturalness as perceived by natives. Further research will therefore focus on the modeling of pauses. Perceptual discrimination, finally, will be tackled following [15] and [16], more specifically between the synthetic versions used in this study.

6. REFERENCES

[1] Abercrombie, D. 1964. Syllable quantity and enclitics in English. In Abercrombie, D., Fry, P., MacCarthy, N., Trim, J. (eds.), *In Honour of Daniel Jones*. London: Longman, 216-222.

[2] André C., Ghio A., Cavé C., Teston B. 2003. PERCEVAL: A computer-driven system for experimentation on auditory and visual perception. *Proceedings of the XVth International Congress of Phonetic Sciences*, 1421-1424.

[3] Auran, C. 2004. *Prosodie et Anaphore dans le Discours en Anglais et en Français: Cohésion et Attribution Référentielle*. PhD dissertation, Université de Provence.

[4] Auran, C., Bouzon, C. 2010. A multi-level approach to speech rate in British English: towards an analysis-by-synthesis method. *Proceedings of the Fifth International Conference on Speech Prosody* 100988, 1-4.

[5] Boersma, P. 2001. Praat. A system for doing phonetics by computer. *Glot International* 5(9/10), 341-345.

[6] Bouzon, C. 2004. *Rythme et Structuration Prosodique en Anglais Britannique Contemporain*. PhD dissertation, Université de Provence.

[7] Bouzon, C., Hirst, D.J. 2004. Isochrony and prosodic structure in British English. *Proceedings of the Second International Conference on Speech Prosody*, 223-226.

[8] Breiman, L., Friedman, J., Olshen, R., Stone, C. 1984. *Classification and Regression Trees*. Monterey, CA: Wadsworth & Brooks.

[9] Grosz, B., Hirschberg, J. 1992. Some intonational characteristics of discourse structure. *Proceedings of the International Conference on Spoken Language Processing* 1, 429-432.

[10] Hirst, D. 2007. A Praat plugin for Momel and Instint with improved algorithms for modelling and coding intonation. *Proceedings of the XVIth International Congress of Phonetic Sciences*, 1233-1236.

[11] Jassem, W. 1952. *Intonation in Conversational English*. Warsaw: Polish Academy of Science.

[12] Koopmans-van Beinum, F.J., van Donzel, M.E. 1996. Discourse structure and its influence on local speech rate. *Proceedings of the Institute of Phonetic Sciences* Amsterdam, 20.

[13] Ladd, D. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[14] Moulines, E., Charbonnier, F. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453-467.

[15] Pfitzinger, H. 1999. Local speech rate perception in German speech. *Proceedings of the XIVth International Congress of Phonetic Sciences*, 893-896.

[16] Quené, H. 2007. On the just noticeable difference for tempo in speech. *Journal of Phonetics* 35, 353-362.

[17] Riley, M. 1992. Tree-based modelling of segmental durations. In Bailly, G., Benoit, C., Sawallis, T.R. (eds.), *Talking Machines: Theories, Models and Designs*. Amsterdam: Elsevier, 265-273.

[18] Smith C. 2004. Topic transitions and durational prosody in reading aloud: Production and modeling. *Speech Communication* 42, 247-270.

[19] Therneau, T.M., Atkinson, B. 2002. *rpart Package, computer software (R library)*, (R port by Ripley B., 2009-08-05).

[20] Trouvain, J. 2003. *Tempo Variation in Speech Production. Implications for Speech Synthesis.* PhD dissertation, Saarbrücken.

[21] Wichmann, A. 2000. *Intonation in Text and Discourse: Beginnings, Middles and Ends.* London: Longman.