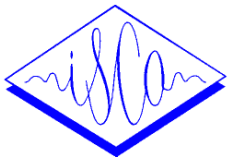

Fourth ISCA ITRW on
Speech Synthesis (SSW-4))
Perthshire, Scotland
August 29 - September 1, 2001

Close shadowing natural vs synthetic speech

Gérard Bailly

Institut de la Communication Parlée
UMR CNRS 5009, INPG & Université Stendhal
46, avenue Félix Viallet, 38031 Grenoble Cedex 1, France
bailly@icp.inpg.fr

Abstract

Close shadowing experiments involving natural and synthetic stimuli are here described. Preliminary results show that speakers are able to follow natural stimuli with an average delay less than 50 ms whereas this delay exceeds 100 ms for stimuli produced by Text-to-speech systems. A complementary experiment shows that this contrast is mainly due to prosody.

1. Introduction

The human ability to shadow speech (i.e., the ability to repeat immediately what is spoken to one) is quite universal: it is independent of native language, language skills, word comprehension and speaker intelligence – many autistic and some mentally retarded people, for instance, echo overheard words (often their only vocal interaction with others) without understanding what they say. It is prelinguistic: eighteen-week-old infants spontaneously copy vocal expressions, provided the accompanying voice matches [10]. Imitation of vowels has been found as young as twelve weeks [11]. It happens quickly: words can be repeated within 250-300 milliseconds, both in normals (during shadowing) [12], and during echolalia by retarded individuals. Moreover, it can be quicker to imitate a syllable than to initiate it. Porter and Lubker [17] found, as they put it, 'simply executing a shift to [o] upon detection of a second vowel in [ao] takes very little longer than does interpreting and executing it as a shadowed response' (p. 599). They suggest for this reason that 'the early phases of speech analysis yield information which is directly convertible to information required for speech production'. A detailed analysis of their results on VCV syllables [16] shows in fact that speakers can trigger the production of the consonant C as soon as the onset of the formant transitions of the preceeding vowel. These results show that speakers may exploit very subttle phonetic details of the driving stimuli to control vocalization.

Speech shadowing seems thus to occur independently of normal speech and provides evidence of a 'privileged' input/output speech loop separate to the other components of the speech system [15]. Neurocognitive research likewise finds evidence of a direct (nonlexical) link between phonological analysis input and motor programming output [14] assessed by the recent discovery of the so-called mirror cells [18]. Complementary results show however that shadowing performance can be influenced by other sources of information about the stimuli including audiovisual presentation [19] or phonological priming [6]. Marslen-Wilson showed [13] that shadowers "were syntactically and semantically analysing the material as they repeated it". He concludes his abstract by stating that "close shadowing provides us with uniquely privileged access to the properties of the system".

This paper presents results from a preliminary experiment comparing performance of natural versus synthetic speech shadowing tasks.

2. Experimental design and procedure

2.1. Material

In all expriments described below, speakers were instructed to repeat back a passage of normal continuous prose: the north wind and the sun (see section 8). The shadowing of the title of the passage is considered as a triggering signal and is excluded of the shadowing performance.

This passage was known in advance by the subjects: the question posed by these experiments was thus not how speakers gather information about *what* to say but *when* to say it. In previous shadowing studies [5, 4, 13] of connected prose where shadowers discovered the message as they heard it, the performance of "distant" shadowers - with average delays between 500 ms to over 2 s - was contrasting with the performance of "close" shadowers able to repeat the speech back at mean latencies of less than 200 ms. Since our experimental design does not involve speech comprehension per se, our subjects are all close shadowers producing average delays of less than 150 ms.

2.2. Experimental setting

The passage was displayed on a computer screen. Stimuli were delivered through earphones with a sound level that was confortable and loud enough to mask their own audio feedback. Duplex stereo recording was used to play the target sound and record simultaneously the earphone signal and shadowed signal: this complex setup was necessary because delays between played and recorded signals were as large as 20 ms despite the triggering mode available on most commerically available sound cards. A simple cross-correlation between the target and earphone signals was performed to determine this delay for each stimuli.

2.3. Instructions

For each stimuli, subjects were first familiarized with the timbre and allowed to shadow the passage in whatever way came naturally to them. They were then asked to shadow as closely as possible. Only two close shadowing trials were allowed but most of the speakers were satisfied with their first trial.

2.4. Measurements

All stimuli were hand labelled using Praat [3]. Since only deletion or omission of phonemes are considered here the alignment was performed favoring labels closest in time using a simple

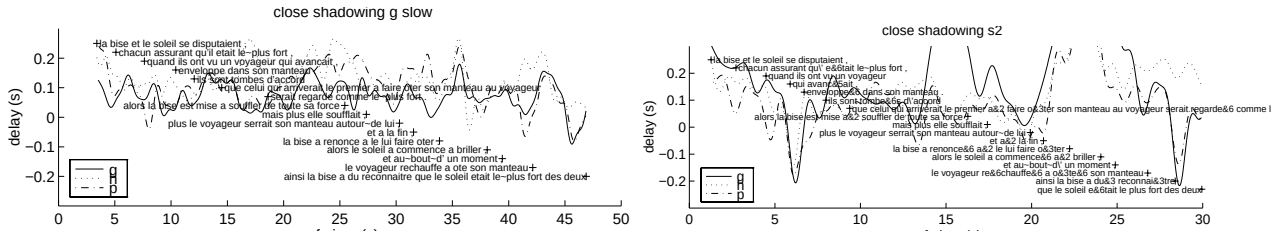

Figure 1: Comparing the Time evolution of the delay for the three close shadowers. Left: for the natural "slow" reference g. Right: for the synthetic reference s2.

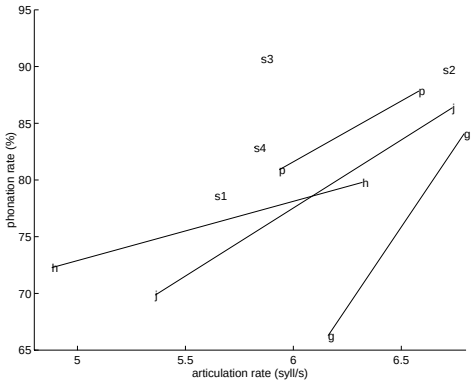

Figure 2: phonation rate as a function of articulation rate for natural target stimuli (the slow and rapid versions of the same speaker are connected with a line) and the four text-to-speech systems considered.

dynamic warping technique. Latency measurements were made at every sound onsets resulting in 401 alignements on average. Examples of such alignements are given Figure 1.

Objective characteristics of alignements that will be commented in the following concern the mean and standard deviations of the latency measurements gathered along each alignement excluding those immediately following target pauses.

2.5. Selecting shadowers

Four speakers (j, h, g and p) initially participated in the shadowing experiment: they will have to shadow their own production (see below) as well as others. All speakers were ICP researchers and could be considered as familiar with speech synthesis and knew eachother well: although our speakers have never conducted before a close shadowing experiment, they can considered as trained subjects and results as optimal performances.

Following the classification proposed by Marslen-Wilson [13], one of these speakers (j) performed as a distant shadower (mean latency > 300 ms) and was thus disgarded as a close shadower.

3. Experiment I. Natural stimuli

3.1. Targets

Four speakers (j, h, g and p) recorded the target stimuli. They were instructed to read aloud the passage with two different styles: (1) as they were reading the story to a child (referenced as the "slow" version), (2) avoiding to pause between full stops (referenced as the "rapid" version).

The figure 2 gives a global view of the subjects performance. Slow versions differ considerably: speaker g has the highest articulation rate but structures its discourse with long pauses, speaker p maintains both high articulation and phonation rates while speaker h slows down both. Rapid versions tend to converge towards an articulation rate of 6.5 syll/s and a phonation rate of 85%.

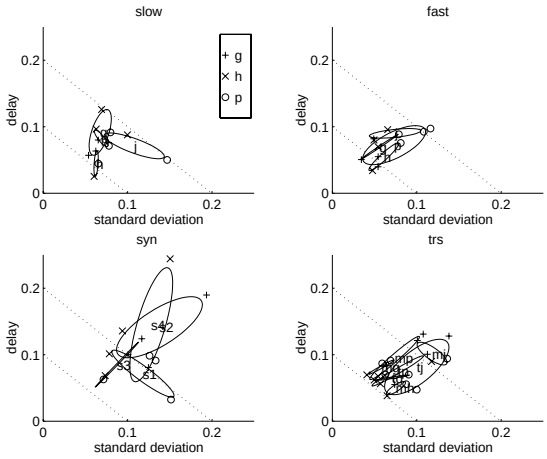

Figure 3: Average and standard deviations of close shadowing latencies according to target.

Table 1: Shadowing latencies: mean and standard deviations (in ms).

speakers	g	h	j	p
slow	68 ± 82	67 ± 91	108 ± 71	62 ± 44
rapid	73 ± 72	56 ± 70	74 ± 90	61 ± 55
systems	s1	s2	s3	s4
	118 ± 71	138 ± 141	87 ± 85	128 ± 145
speakers	g	h	j	p
mbrola	58 ± 81	75 ± 50	118 ± 102	74 ± 95
tdpsola	71 ± 65		100 ± 81	78 ± 76

3.2. Results & discussion

Global results are summarized in figure 3 and table 1: all mean latencies lay below 100 ms. This delay is far below the average results obtained by previous studies either considering isolated

syllabes known in advance [16] or connected prose the content of which speakers discovered when shadowing [13]. This difference could easely be explained by the fact that speakers could exploit here far more top-down information for predicting the temporal structure of the speech to come: as the text is known in advance, congruency between prosody and text and information structure can be fully exploited while still exploiting general properties of prosodic structures such as long-term coherence and predictibility [9, 1].

Reading style does not seems to influence the shadowing performance: despite large differences in speaking rates and phrasing strategies, all natural references are shadowed around 70 ms with the exception of the slow version of speaker p who employs an unusual reading style for telling stories to children!

Surprisingly speaker j - who was disgarded as a distant shadower - was also the most difficult speaker to shadow.

4. Experiment II. Synthetic stimuli

4.1. Targets

The passage was synthetized by four French text-to-speech systems available on the web. Two of them did not allow the synthesis of a complete paragraph and the passage had to be processed sentence by sentence. In this latter case, we set an ad hoc rule for generating pauses between sentences by imposing an average phonation rate of 80% i.e. between the duration of each pause and the duration of its adjacent sentences. The minimum pause duration was 250 ms. These synthetic stimuli were collected during July 2000. They are referenced in the following as stimuli s1, s2, s3 and s4.

All these sytems use concatenative synthesis with different male voices and could be considered as representing the state of the art of French synthesis.

4.2. Results & comments

Global results are summarized in figure 3 and table 1: systems s1 and s3 reach performance close to natural targets whereas systems s2 and s4 resulted in worst and more scattered performance. s1 is a commercial system which results from a long-term research effort from both industrial and academic institutions and it is not surprinsing to have also this system ranked subjectively by close shadowers as producing the easiest stimuli to shadow. Results obtained by s3 are more intriguing: its phonation rate is too high compared to its articulation rate (see figure 2) and this system is ranked subjectively by close shadowers as producing the most difficult stimuli to shadow. Closer inspection of the rhythmic structure of s3 stimuli shows that s3 produces the smallest standard deviation of syllabic durations: isochronous syllables seems thus to be easy to shadow but at the expense of a larger cognitive effort. This poor rhythmic structure should therefore handicap shadowers when the message content is not known in advance and affects comprehension such as suggested by Marslen-Wilson [13].

5. Experiment III. Copy synthesis

We question here what causes the worse performance of the shadowers in case of synthetic speech: is this caused (a) by the poor segmental quality of the signals in which the close shadowers do not find good or sufficiently clear low level acoustic cues (such as formant transitions as suggested by Porter and colleagues [16]) for triggering their responses or (b) by an inappropriate rhythmical - prosodic - organization of these acoustic cues.

5.1. Targets

We give here results of a last close shadowing experiment using synthetic stimuli produced by feeding two different concatenative sysnthesis systems - one using MBROLA [7] and one using TDPSOLA [2] - with the segmental durations and the appropriate stylization of the melody of the “slow” versions of experiment I. Seven copy synthesis targets were computed: 6 copy synthesis of target uttered by speakers j, g and p using MBROLA with the male voice fr1 - referenced as stimuli mj, mg and mp - and TDPSOLA with the ICP male segment database - referenced as stimuli tj, tg and tp - and 1 copy synthesis of the “slow” version of the female speaker h using MBROLA with the female voice fr3 - referenced as stimulus th -.

5.2. Results & discussion

Global results are summarized at the bottom of figure 3 and table 1: all mean latencies lay below 100 ms. These results are very closed to those obtained in experiment I. We do find the relative difficulty for speakers to shadow stimuli from speaker j whenether synthetized by MBROLA or TDPSOLA systems used in this paper.

This third experiment shows that most of the increase of latencies observed for synthetic stimuli should be accounted to the impoverished prosody they are able to generate from raw text. On the contrary it shows that concatenative synthesis produces a signal that is rich enough to anchor properly our perception of the rhythmic structure of the stimuli.

6. Conclusions

The preliminary close shadowing experiments conducted here do not make use of a large panel of subjects and all subjects were familiar to speech synthesis. We plan to investigate the performance of more naive subjects in front of more impoverished signals and without textual guidance. These preliminary experiments show however that fine objective distinctions could be made even when all variables should contribute to reduce variability.

Although speakers were not instructed specifically to mimic the speeches as closely as possible, a small but significant tendency to mimick fundamental frequency (F0) targets can be seen in Figure 4. Close shadowing and mimicry exhibit however inverse timing and F0 performance: impersonators [8] succeed quite well in attaining both global and local F0 targets and also global speech rate whereas local deviations may rise up to 1.5 s. This is certainly not the case here. It will be however interesting in the near future to investigate the consequences of such an additionnal instruction on the close shadowing performance of the speakers.

Finally we should emphasize that these experiments have been made possible because of the availability of text-to-speech servers on the web. Although such experiments deliver an instantaneous photograph of a system which is always “under construction”, these systems offer a unique way of gathering and studying the variability of synthetic speakers.

7. Acknowlegments

Two master students, Laurie Champion and Anne Tripier, worked hard on the two first experiments.

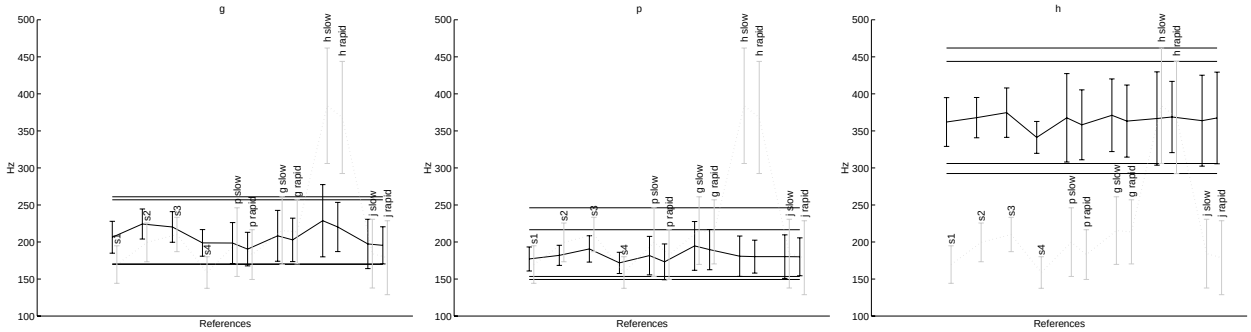

Figure 4: Comparing F0 means and standard deviations of targets and subjects' responses.

8. Appendix: reading material

La bise et le soleil.
La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort, quand ils ont vu un voyageur qui s'avançait, enveloppé dans son manteau.
Ils sont tombés d'accord que celui qui arriverait le premier à faire ôter son manteau au voyageur serait regardé comme le plus fort.
Alors la bise s'est mise à souffler de toute sa force, mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui, et à la fin, la bise a renoncé à le lui faire ôter.
Alors le soleil a commencé à briller, et au bout d'un moment, le voyageur réchauffé a ôté son manteau.
Ainsi la bise a dû reconnaître que le soleil était le plus fort des deux."

9. References

[1] Aubergé, V., Grépillat, T., and Rilliard, A. Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 2, pages 871–874, Rhodes - Greece, 1997.

[2] Bailly, G., Barbe, T., and Wang, H. Automatic labelling of large prosodic databases: tools, methodology and links with a text-to-speech system. In Bailly, G. and Benoît, C., editors, *Talking Machines: Theories, Models and Designs*, pages 323–333. Elsevier B.V., 1992.

[3] Boersma, P. and Weenink, D. Praat, a system for doing phonetics by computer, version 3.4. Institute of Phonetic Sciences of the University of Amsterdam, Report 132. 182 pages.

[4] Carey, P. Verbal retention after shadowing and after listening. *Perception and Psychopysics*, 9:79–83, 1971.

[5] Chistovich, L.A., Aliakrinskii, V., and Abulian, V. Time delays in speech repetition. *Voprosy Psikhologii*, 1:114–119, 1960.

[6] Dumay, N. and Radeau, M. Rime and syllabic effects in phonological priming between french spoken words. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 4, pages 2191–2194, 1997.

[7] Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and van der Vrecken, O. The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of the International Conference on Speech and Language Processing*, volume 3, pages 1393–1396, Philadelphia - USA, 1996.

[8] Eriksson, A. and Wretling, P. How flexible is the human voice? a case study of mimicry. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 1043–1046, Rhodes - Greece, 1997.

[9] Grosjean, F. How long is the sentence? prediction and prosody in the on-line processing of language. *Linguistica*, 21:501–529, 1983.

[10] Kuhl, P.K. and Meltzoff, A.N. The bimodal perception of speech in infancy. *Science*, 218:1138–1141, 1982.

[11] Kuhl, P.K. and Meltzoff, A.N. Infant vocalizations in response to speech: Vocal imitation and developmental change. *Journal of the Acoustical Society of America*, 100:2425–2438, 1996.

[12] Marslen-Wilson, W. Linguistic structure and speech shadowing at very short latencies. *Nature*, 244:522–523, 1973.

[13] Marslen-Wilson, W. Speech shadowing and speech comprehension. *Speech Communication*, 4:55–73, 1985.

[14] McCarthy, R. and Warrington, E.K. A two-route model of speech production: Evidence from aphasia. *Brain*, 107:463–485, 1984.

[15] McLeod, P. and Posner, M.I. Privileged loops from percept to act. In Bouma, H. and Bouwhuis, D., editors, *Attention and performance X*, pages 55–66. Lawrence Erlbaum Associates, Mahwah, NJ - USA, 1984.

[16] Porter, R.J. and Castellanos, F.X. Speech-production measures of speech perception: rapid shadowing of VCV syllables. *Journal of the Acoustical Society of America*, 67(4):1349–1356, 1980.

[17] Porter, R.J. and Lubker, J.F. Rapid reproduction of vowel-vowel sequences: evidence for a fast and direct acoustic-motoric linkage in speech. *Journal of Speech and Hearing Research*, 23:593–602, 1980.

[18] Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141, 1996.

[19] Vitkovitch, M. and Barber, P. Effect of video frame rate on shadowing. *Journal of Speech and Hearing Research*, 37:1204–1210, 1994.