INTERNATIONAL JOURNAL OF SPEECH TECHNOLOGY 6, 11–19, 2003

Close Shadowing Natural Versus Synthetic Speech

G. BAILLY
Institut de la Communication Parlée, UMR CNRS n°5009, INPG/Univ. Stendhal, 46, av. Félix Viallet, 38031 Grenoble CEDEX France
bailly@icp.inpg.fr

Abstract. Close shadowing experiments involving natural and synthetic stimuli are described. Preliminary results show that speakers are able to follow natural stimuli with an average delay of 70 ms whereas this delay typically exceeds 100 ms for stimuli produced by text-to-speech systems. A complementary experiment shows that this contrast is mainly due to the inappropriate or impoverished prosody generated by actual text-to-speech systems.

Keywords: evaluation, text-to-speech synthesis, prosody, close shadowing

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 (Kuhl and Meltzoff, 1982). Imitation of vowels has been found as early as twelve weeks (Kuhl and Meltzoff, 1996). It happens quickly. Words can be repeated within 250–300 milliseconds, both during shadowing by normal people (Marslen-Wilson, 1973) and during echolalia by mentally retarded individuals (Fay and Coleman, 1977; Schneider, 1938). Moreover, it can be quicker to imitate a syllable than to initiate it. Porter and Lubker (1980) 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 (Porter and Castellanos, 1980) shows in fact that speakers can trigger the production of the consonant C as soon as the onset of the formant transitions in the preceding vowel. These results show that speakers may exploit very subtle 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 independent of the other components of the speech system (McLeod and Posner, 1984). Neurocognitive research likewise finds evidence of a direct (non-lexical) link between phonological analysis input and motor programming output (McCarthy and Warrington, 1984) supported by the recent discovery of the so-called mirror cells (Rizzolatti et al., 1996). Complementary results show, however, that shadowing performance can be influenced by other sources of information about the stimuli including audiovisual presentation (Vitkovitch and Barber, 1994) or phonological priming (Dumay and Radeau, 1997). Marslen-Wilson (1985) showed that shadowers "were syntactically and semantically analyzing the material as they repeated it". He concludes "*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. We will investigate if shadowing performance is influenced by the impoverished phonetic (both segmental and prosodic) structure of synthetic stimuli.

2. Experimental Design and Procedure

2.1. Material

In all experiments described below, speakers were instructed to shadow a passage of normal continuous prose—the north wind and the sun (see Appendix)—as close as possible to a target reading. The shadowing of the title of the passage is considered as a triggering signal and is excluded from shadowing analysis.

Target readings have been obtained by the reading of the same passage by human speakers (including themselves), by a text-to-speech synthesizer and by copy synthesis. These experiments are described respectively in Sections 3, 4 and 5.

In the following, we will characterize (and eventually rank) each target speaker or system according to the performance of four subjects that will shadow the different target readings. The shadowing performance is evaluated by computing the time-varying delay between each target reading and the shadowed responses (see Section 2.4).

This passage is known in advance by all the subjects. The question posed by these experiments is thus not how quickly speakers gather information about *what* to say but only *when* to say it. In previous shadowing studies (Carey, 1971; Chistovich et al., 1960; Marslen-Wilson, 1985) of connected prose, shadowers discovered the message as they heard it. Typically, two types of shadowers are identified: 'distant' shadowers, with average delays between 500 ms to over 2 s, and '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, most of our subjects are close shadowers in that sense, as they produce average delays of less than 150 ms.

2.2. Experimental Setting

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

2.3. Instructions

For each target stimulus, subjects were first familiarized with the speaker's or synthesis system's characteristics. They listened to the target stimulus and then shadowed the passage in whatever way came naturally to them. They were then asked to shadow as closely as possible. Although two close shadowing trials were allowed, most of the speakers were satisfied with their first trial. Only the best performance—according to the subject—was retained for further analysis.

2.4. Measurements

All target and shadowed stimuli are first automatically aligned with a normative transcription including optional pauses using Viterbi decoding of phoneme-sized hidden Markov models trained by HTK (Young, 1992), then the labels and their boundaries are hand-corrected using the signal editor Praat (Boersma and Weenink, 1996). We then align target and shadowed allophones using a simple dynamic warping technique. Latency measurements are made for each target sound onset resulting in 401 alignments on average. Examples of these measurements for two different target stimuli are given in Fig. 1.

Objective characteristics of these latencies that will be discussed below, namely the mean and standard deviations of the latency measurements gathered along each alignment path excluding those adjacent to the target pauses of target stimuli. In fact, phoneme boundaries are most imprecise around pauses for diverse reasons including energy drops for final allophones, aspiration of initial consonants, lack of acoustic onset for initial stops, etc.

2.5. Selecting Shadowers

Four speakers (g, h, j and p) initially participated in the shadowing experiment. They had 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. They knew each other well. Although the speakers had never conducted before a close shadowing experiment, they can be considered

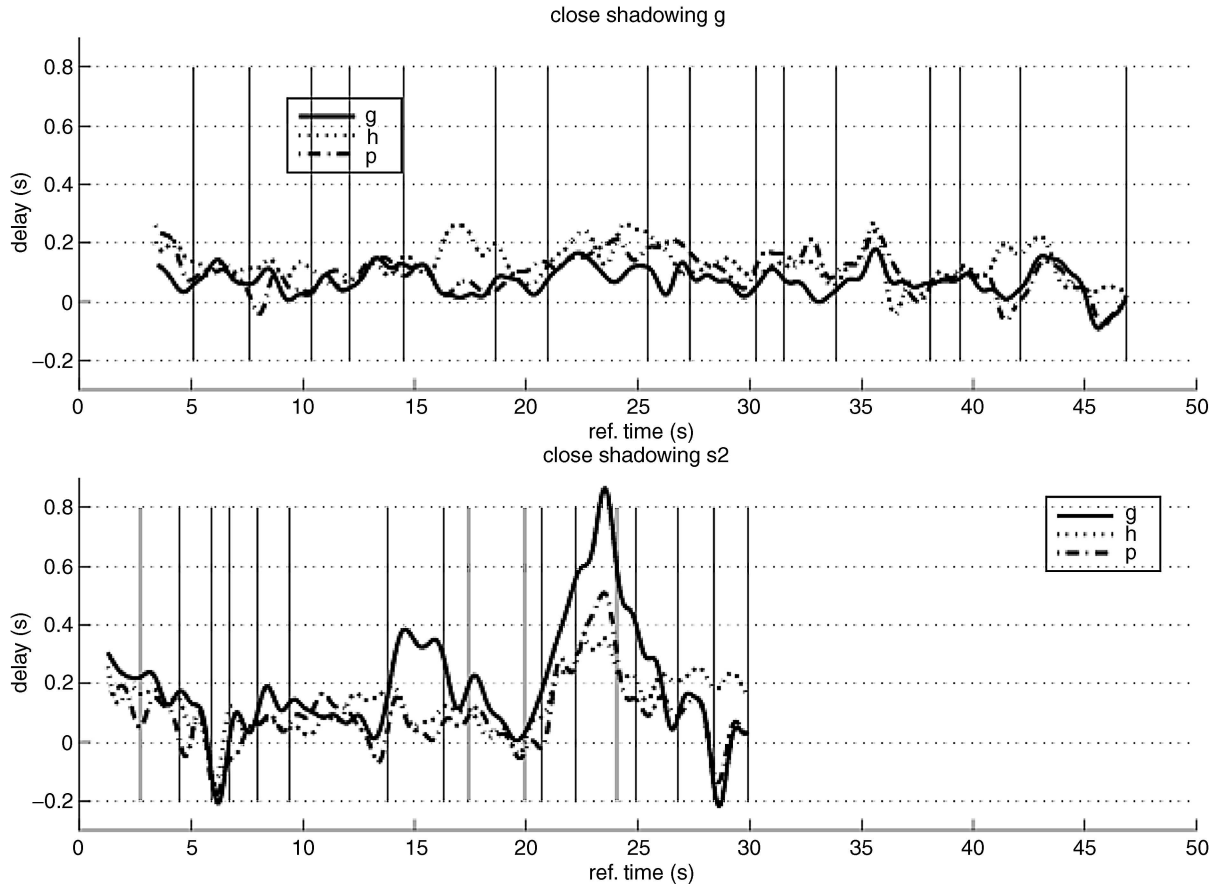

Figure 1. Comparing the time evolution of the delays between a target stimulus and shadowed responses from the three close shadowers g, h and p (see Section 2.5). Above: for the natural 'slow' target from speaker g (see Section 3.1). Bottom: for the synthetic target computed by the text-to-speech system s2 (see Section 4.1). Vertical bars indicate the beginnings of phonation after silent pauses for the target stimulus. Target stimuli have quite different lengths: as seen on Fig. 2, g has both lower articulation rate and phonation rate than s2.

as trained subjects and the results as close to optimal performances.

Following the classification proposed by Marslen-Wilson (1985), one of these speakers (j) performed as a distant shadower (mean latency >300 ms) and was thus discarded as a close shadower.

3. Experiment I. Natural Stimuli

3.1. Targets

The four speakers (g, h, j and p) recorded the target stimuli. They were instructed to read aloud the passage with two different styles: (1) as if they were reading the story to a child (referenced as the 'slow' version), (2) not pausing between full stops (referenced as the 'rapid' version). We expect that the first instruction elicits hyperarticulated speech, short intonation phrases and rather long pauses whereas the second instruction elicits more hypoarticulated speech, longer intonation phrases and shorter pauses. We will then be able to examine if continuous signals favor lower shadowing latencies and if shadowers can effectively predict with the same reliability how long are the pauses as they do with the length of an utterance (Grosjean, 1983; Grosjean and Hirt, 1996).

Figure 2 gives a global view of the speakers' performance. We characterize each paragraph reading by two parameters: the phonation rate and the articulation rate. The phonation rate is the quotient between the duration of effective phonation and the total duration of the reading. The articulation rate is the average number of syllables uttered during a second of effective phonation.

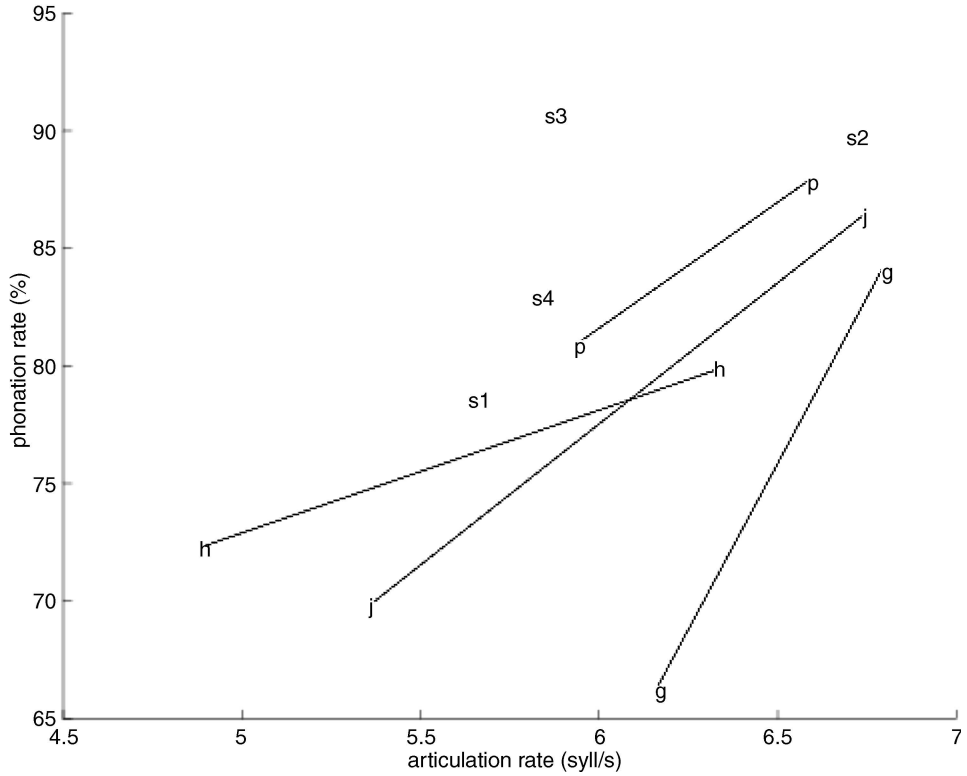

Figure 2. Phonation rate as a function of articulation rate for the natural target stimuli (the slow and rapid versions of the same speaker are connected with a line) and the outputs of the four text-to-speech systems considered.

Slow versions differ considerably: speaker g has the highest articulation rate but produces 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 syllables per second and a phonation rate of 85%.

3.2. Results and Discussion

Global results are summarized in Fig. 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 syllables known in advance (Porter and Castellanos, 1980) or connected prose, the content of which speakers discovered when shadowing (Marslen-Wilson, 1985).

This difference could easily 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 predictability (Aubergé et al., 1997; Grosjean, 1983). This rhythmical predictability may result from low-level 'biological' rhythmical constraints such as provided by the jaw cyclic attractor that regulates the succession of consonants and vowels in most of the world's languages. (Initially put forward by

Table 1. Shadowing latencies: mean and standard deviations (in ms). The copy synthesis system TDPSOLA has no female voice and could not reproduce the prosody of the speaker h.

Speaker	g	h	j	p
Slow	68 ± 82	67 ± 91	108 ± 71	62 ± 44
Rapid	73 ± 72	56 ± 70	74 ± 90	61 ± 55
Text-to-speech system	s1	s2	s3	s4
	118 ± 71	138 ± 141	87 ± 85	128 ± 145
Prosody from speaker	g	h	j	p
MBROLA	58 ± 81	75 ± 50	118 ± 102	74 ± 95
TDPSOLA	71 ± 65		100 ± 81	78 ± 76

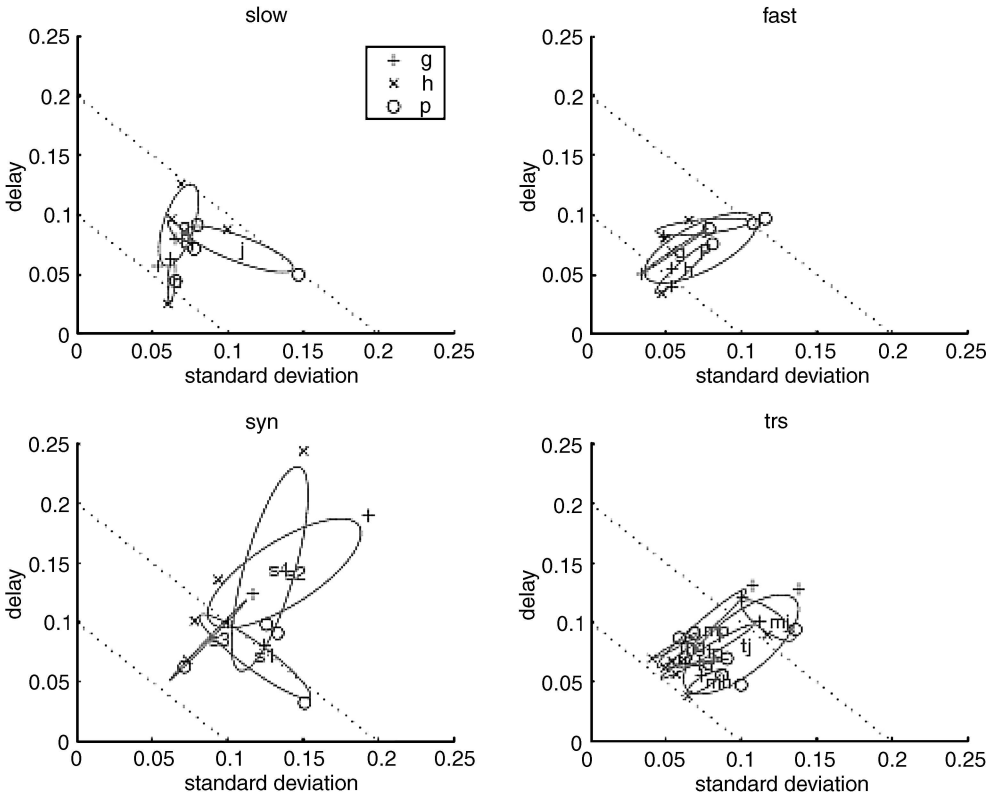

Figure 3. Average and standard deviations of close shadowing latencies computed for each (target stimulus, shadower). Dispersion ellipses characterize each target 'system'. From left to right, top to bottom: shadowing characteristics are displayed respectively for the slow and fast versions of the passage uttered by our four human speaker (see Section 3.1), for the synthetic targets computed by the four text-to-speech systems (see Section 4.1) and the seven close copy synthesis target stimuli (see Section 5.1).

Stetson (1905), this is known as the "pure frame" hypothesis (MacNeilage, 1998) and explains the formation of the first words by a pure jaw oscillation with passive lips and tongue. The natural—most comfortable—jaw oscillation frequency is 5–7 Hz, close to the average articulation rate.) An additional factor may be language and speaker's specific rhythmical patterns. This notion of *rhythmical expectation* is also proposed for music perception (Auxiette and Gérard, 1992; Jones and Boltz, 1989; Schmuckler, 1989).

The mean and standard deviation of observed latencies are highly correlated: the smaller is the latency, the more constant it remains. On the contrary when the latency increases, the more variation of the latency we observe. This could be due partly to the buffering and cushioning effects of longer latency that cause the speaker not to react promptly to unexpected perturbations of the rhythm he/she predicted for the upcoming target stimuli from what he/she already listened to. Another explanation is that the prediction itself may be simply incorrect because of the unusual or incoherent prosodic structure of the target.

Subjects do not shadow their own speech stimuli with significantly less latency. Reading style does not seem 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 discarded as a distant shadower—was also the most difficult speaker to shadow.

4. Experiment II. Synthetic Stimuli

4.1. Targets

The passage was synthesized 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 systems use concatenative synthesis with different male voices and could be considered as representing the state of the art of French text-to-speech synthesis.

4.2. Results and Comments

Global results are summarized in Fig. 3 and Table 1. Systems s1 and s3 reach performance close to natural targets whereas the performance of systems s2 and s4 is worst and more scattered. System s1 is a commercial product which results from a long-term research effort from both industrial and academic institutions and it is not surprising that this system was ranked subjectively by close shadowers as producing the easiest stimuli to shadow. Results obtained with s3 are intriguing: its phonation rate is too high compared to its articulation rate (see Fig. 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 seem 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 as suggested by Marslen-Wilson (1985). This is, however, quite speculative and should be investigated in the near future.

5. Experiment III. Copy Synthesis

We question here what causes the worse performance of the shadowers in the case of synthetic speech. Is this caused by (a) 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 (1980)) for triggering their responses or (b) an inappropriate rhythmical—prosodic—organization of these acoustic cues? These causes are probably not mutually exclusive.

5.1. Targets

We give here results of a last close shadowing experiment using synthetic stimuli produced by feeding two different concatenative synthesis systems—one using MBROLA (Dutoit et al., 1996) and one using a customized implementation (Bailly et al., 1990) of TDPSOLA (Charpentier and Moulines, 1990)—with the segmental durations and the appropriate stylization of the melody of the 'slow"' versions of Experiment I. Seven copy synthesis targets were computed:

- Three copy syntheses of the slow versions of speakers g, j and p using MBROLA with the male voice fr1 (referenced respectively as stimuli mj, mg and mp).
- Same as above but using TDPSOLA with the ICP male segment database (referenced respectively as stimuli tj, tg and tp).
- One copy synthesis of the slow version of the female speaker h using MBROLA with the female voice fr3 (referenced as stimulus mh).

5.2. Results and Discussion

Global results are summarized at the bottom of Fig. 3 and Table 1: all mean latencies lie below 120 ms. These results are very close to those obtained in Experiment I. Speakers also report the same difficulty in shadowing stimuli from speaker j when synthesized by either the MBROLA or the TDPSOLA systems used in this paper.

This third experiment shows that most of the increase of latencies observed for synthetic stimuli should be attributed to the impoverished prosody which current synthesis systems 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. Shadowing and Imitation

Although speakers were not instructed specifically to mimic the speech as closely as possible, a small but significant tendency to mimic fundamental frequency ($F0$) targets can be seen in Fig. 4. Close shadowing and

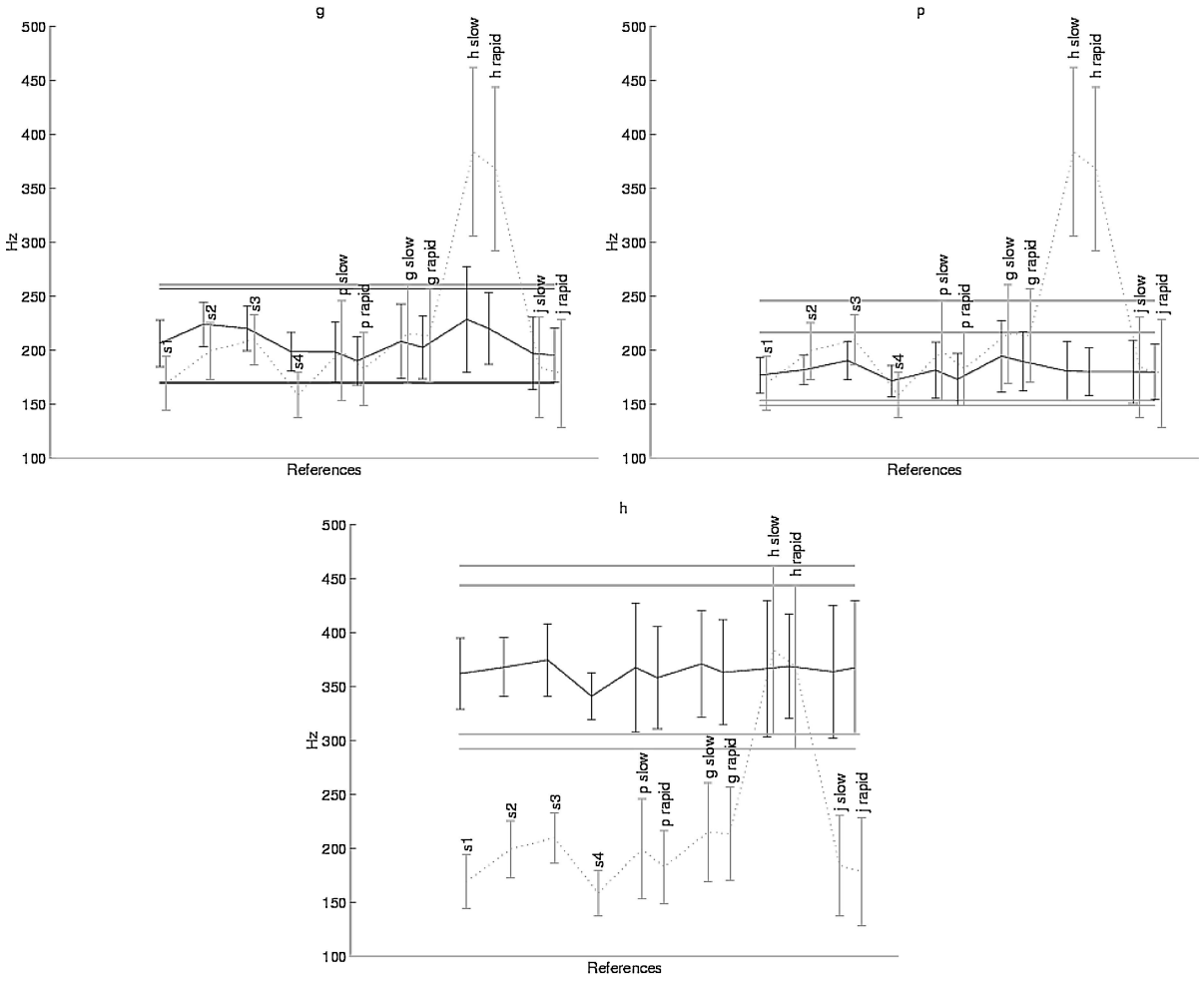

Figure 4. Comparing $F0$ means and standard deviations of targets (light gray) and subjects' responses (plain lines). Horizontal lines are ± one standard deviation away from the mean of speaker's stimuli and shadows. The subjects clearly stay within their normal frequency range.

mimicry exhibit inverse timing and $F0$ performance: impersonators (Eriksson and Wretling, 1997) 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 interesting in the near future to investigate the consequences of such an additional instruction—imitate target timbre and intonation—on the close shadowing performance of the speakers.

7. Conclusions

This series of experiments shows that the close shadowing paradigm can be considered as a relevant tool for evaluating synthetic speech, especially for evaluating the 'naturalness' of synthetic prosody. The adequacy of computed prosodic parameters for encoding discourse structure is supposed to be directly reflected in the amplitude of the latency between the synthetic stimuli and collected shadows.

Close shadowing experiments such as proposed and tested here aim at revealing subtle differences between the *online* processing of synthesized speech and human speech and even differences between synthesis techniques and strategies. A more detailed analysis of temporal structures of shadowing latencies should be conducted. A preliminary analysis has evidenced for example that an appropriate pause duration generates large excursions of the target/shadow latency (both negative and positive—see Fig. 1). This has to be interpreted in the framework of a rhythmical expectation

paradigm (such as evidenced by Grosjean (1983) and Grosjean and Hirt (1996)) that includes pause location and duration.

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

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 speech.

Appendix

The read corpus used in this experiment was displayed (and also delivered to text-to-speech synthesizers) as follows:

« 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. »

Acknowledgments

Two masters students, Laurie Champion and Anne Tripier, contributed to the two first experiments. This work benefits from fruitful discussions with Plinio Barbosa and Fred Cummins. This paper greatly benefits from the thoughtful comments of Eric Bateson and two other anonymous reviewers.

References

Aubergé, V., Grépillat, T., and Rilliard, A. (1997). Can we perceive attitudes before the end of sentences? The gating paradigm for prosodic contours. *Proceedings of the European Conference on Speech Communication and Technology*. Rhodes, Greece, pp. 871–874.

Auxiette, C. and Gérard, C. (1992). Perceptual and motor determinants in the synchronization of music and speech. *Fourth International Workshop on Rhythm Perception and Production*. Bourges, France, pp. 59–64.

Bailly, G., Barbe, T., and Wang, H. (1990). Automatic labelling of large prosodic databases: Tools, methodology and links with a text-to-speech system. *ETRW Workshop on Speech Synthesis*. Autrans, France, pp. 201–204.

Boersma, P. and Weenink, D. (1996). Praat, a system for doing phonetics by computer, version 3.4, Institute of Phonetic Sciences of the University of Amsterdam, Report 132. 182 pages.

Carey, P.W. (1971). Verbal retention after shadowing and after listening. *Perception and Psychopysics*, *9*:79–83.

Charpentier, F. and Moulines, E. (1990). Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, *9*(5/6):453–467.

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

Dumay, N. and Radeau, M. (1997). Rime and syllabic effects in phonological priming between French spoken words. *Proceedings of the European Conference on Speech Communication and Technology*, pp. 2191–2194.

Dutoit, T., Pagel, V., Pierret, N., Bataille, F., and Vrecken, O.v.d. (1996). The MBROLA project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. *Proceedings of the International Conference on Speech and Language Processing*. Philadelphia, USA, pp. 1393–1396.

Eriksson, A. and Wretling, P. (1997). How flexible is the human voice? A case study of mimicry. *Proceedings of the European Conference on Speech Communication and Technology*. Rhodes, Greece, pp. 1043–1046.

Fay, W.H. and Coleman, R.O. (1977). A human sound transducer/reproducer: Temporal capabilities of a profoundly echolatic child. *Brain and Language*, *4*:396–402.

Grosjean, F. (1983). How long is the sentence? Prediction and prosody in the on-line processing of language. *Linguistica*, *21*:501–529.

Grosjean, F. and Hirt, C. (1996). Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language and Cognitive Processes*, *11*(1):107–134.

Jones, M.R. and Boltz, M.G. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*:459–491.

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

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

MacNeilage, P. (1998). The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences*, *21*(4):499–548.

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

Marslen-Wilson, W. (1985). Speech shadowing and speech comprehension. *Speech Communication*, *4*:55–73.

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

McLeod, P. and Posner, M.I. (1984). Privileged loops from percept to act. In H. Bouma and D. Bouwhuis (Eds.), *Attention and performance X*. Lawrence Erlbaum Associates: Mahwah, NJ, USA, pp. 55–66.

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

Porter, R.J. and Lubker, J.F. (1980). 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.

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

Schmuckler, M. (1989). Expectation in music: Investigation of melodic and harmonic processes. *Music Perception*, *7*:109–150.

Schneider, D.E. (1938). The clinical syndromes of echolalia, echopraxia, grasping and sucking. *Journal of Nervous and Mental Disease*, *88*(18–35):200–216.

Stetson, R.H. (1905). Motor theory of rhythm and discrete succession I and II. *Psychological Review*, *12*:250–269, 293–335.

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

Young, S.J. (1992). *HTK: Hidden Markov Model Toolkit V1.3. Reference Manual*. Cambridge University Engineering Department.