Proceedings of the Institute of Acoustics

USING SYNCHRONOUS SPEECH TO MINIMIZE VARIABILITY

Fred Cummins Department of Computer Science, University College Dublin, Dublin 4, Ireland
Deb Roy 20 Ames Street, Rm. E15-384C, MIT Media Lab, Cambridge, MA 01242, USA

1 INTRODUCTION

Many experimental methods seek to minimize phonetic variation, hoping to spare linguistic information and eliminate the paralinguistic, the idiosyncratic and the "merely" expressive [3]. A conventional technique (phonetic vice clamps) is the use of a frame such as "say X again" , familiar from many studies. Typically, the underlying assumption is that utterance-to-utterance variability stems from the combination of two independent sources. The first is the linguistic specification of an item (including phonological specification and coarticulatory effects), and the second is a set of non-linguistic factors associated with expressiveness, affect, communicative urgency, channel characteristics and the like. As the first set is of particular interest to the experimental phonetician or the laboratory phonologist, methods are employed which serve to reduce the latter while sparing the former.

We are currently investigating a new method of reducing variability which serves precisely this function. The method is designed to tap a speaker's own unconscious estimate of what is expendable and what essential in speech, without requiring explicit knowledge of speech variables, without extensive practice and without undue sacrifice of "naturalness". This is achieved by the relatively simple expedient of having subjects read prepared texts in synchrony with one another. Typically, two subjects read a short text aloud together, each endeavoring to maintain synchrony with the other. We call speech elicited in this fashion *Synchronous Speech*.

Our initial experiments with Synchronous Speech (hereafter, SS) have served to illustrate the following features:

- Subjects can read texts together with a high degree of synchrony
- Little or no practice is required to achieve this goal
- When reading in synchrony, subjects exhibit a high level of agreement on matters which display considerable variability under other circumstances
- The resulting speech is expressively neutral, while linguistic information does not appear to be lost.

A priori consideration of the SS elicitation condition suggests that the goal of speaking in tight synchrony with another speaker can be achieved only if the speakers manage to each make their speech

temporally predictable for the other. This in turn means agreeing on common timing patterns for speech, and thus suggests that speakers must exploit their shared knowledge of speech timing and dispense with idiosyncratic and unpredictable elements in their speech. There is thus every reason to believe that if subjects can achieve relatively tight synchrony, a reduction in inter-subject variability will result. A parallel situation exists in music: ensemble playing is much less variable than the performance of a soloist [4]. Requiring musicians to play together has the effect of restricting their idiosyncratic variation, while enhancing their common interpretation of a score.

Speaking in approximate synchrony with other speakers is familiar from tasks such as praying, chanting, reciting oaths, etc. A feature of these situations is that the texts are well practiced, and usually have a highly stylized prosody. For example, the Oath of Allegiance, as recited by American schoolchildren differs markedly from a reading by one unfamiliar with the text. However, we have found that novel texts can also be read in synchrony, and an informal trial may serve to convince the reader that this task is not particularly demanding to the native speaker.

We believe SS will prove to be a useful tool in the experimental arsenal of the phonologist and phonetician, and may prove to be of use in the development of speech technologies as well. In this paper, we will illustrate the use of this technique to reduce variability. In a small experiment, we look specifically at pause placement while reading complex sentences, and demonstrate that SS serves to reduce variability across speakers. We then consider a possible exploitation of SS in a concatenative speech synthesis application.

2 EXPERIMENT 1: SYNCHRONOUS SPEECH

2.1 Methods

Four subjects (2 m, 2 f, age 20–35) participated. All were from the area around Dublin, Ireland. Readings of the first part of the Rainbow Text (see Table 1) were obtained in three conditions. In the 'solo' condition, subjects first practiced reading the text aloud, after which 12 recordings were obtained, without any further constraints on speaking style or rate. In the 'recording' condition, each speaker attempted to read the text in synchrony with a recording (from the first session) of one of the other speakers. 12 trials per subject were obtained (4 target recordings taken randomly from each of the 3 other subjects). Finally, in the 'synchronous' condition, each subject-pair read the text 4 times in synchrony. In this latter condition, subjects were seated comfortably next to one another. Each wore a head-mounted near-field microphone (Shure WH20), and recordings were made onto the right and left channels of a single stereo file. Subjects were free to look at one another throughout.

2.2 Results

We first compare the solo condition with the synchronous condition.

[When the sunlight strikes raindrops in the air they act like a prism and form a rainbow]
[The *rain*bow is a division of white light into many beautiful *col*ors]
[*These* take the shape of a long round arch with its path high above, and its two ends apparently *be*yond the horizon]
[*There* is, according to legend, a boiling pot of gold at one *end*]
[*Peo*ple look, but no one ever finds *it*]
[*When* a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rain*bow*]

Table 1: Canonical division of the first paragraph of the rainbow text into 6 phrases. Vowel onsets in italicized syllables formed the basis of measurements of synchrony reported below.

subject	solo	recording	synchronous
F1	97 (5.3)	77 (5.8)	68 (7.8)
F2	70 (6.5)	49 (7.8)	45 (7.5)
M1	58 (5.2)	48 (4.9)	32 (3.2)
M2	41 (8.5)	46 (13.3)	30 (5.8)

Table 2: Mean (s.d.) range within which 90% of measured F_0 values fell within a trial.

2.2.1 Phrasing

Table 1 divides the text into 6 distinct phrases. In the 'synchronous' condition, pauses (silence of more than 200 ms) are present at these phrase edges without exception. Pauses at other points (such as major syntactic edges within these phrases) occur 4 times in 24 paired readings. By contrast, pauses occur at other points 48 times in the 48 readings in the solo condition. Pauses are absent at these phrase edges 4 times (one speaker only). Thus speakers display almost complete agreement on pause placement in the synchronous condition, but often add additional pauses when reading alone.

2.2.2 Rate

Three speakers show a longer utterance duration in the synchronous condition, and one a shorter. The ratio of pause (silence longer than 200 ms) to speech is larger in the synchronous condition for two of the four speakers, and smaller for the other two. Thus there is no consistent effect of condition on either speech rate or articulation rate.

2.2.3 Pitch range

Pitch range (operationally defined as the range within which 90% of measured pitch values lie) is reduced in both the recording and the synchronous condition. Table 2 gives means and standard deviations for each speaker and condition. Pitch range is greatly reduced in the synchronous condition. Although we believe synchronous speech will allow empirical progress in several vexing questions related to intonation, we defer such analysis to future experiments.

We now turn our attention to measurement of synchrony in the recording and synchronous conditions.

2.2.4 Synchrony

For each phrase but the first, the magnitude of the lag between corresponding vowel onsets at the start and at the end of the phrase was measured. The syllables used are given in italics in Table 1. The data are plotted in Figure 1. In each case, synchrony is greater in the synchronous condition than in the recording condition. A Wilcoxon signed rank test confirms the effect of position within each condition and a Wilcoxon rank sum test compares corresponding positions across conditions (all p-values < 0.001). It is especially interesting to see that synchrony at the beginning of phrases in the synchronous condition appears to be as good as at the end of phrases in the recording condition. Subjects in the synchronous condition thus have a much easier time in predicting when speech will resume after a pause at a major phrase edge.

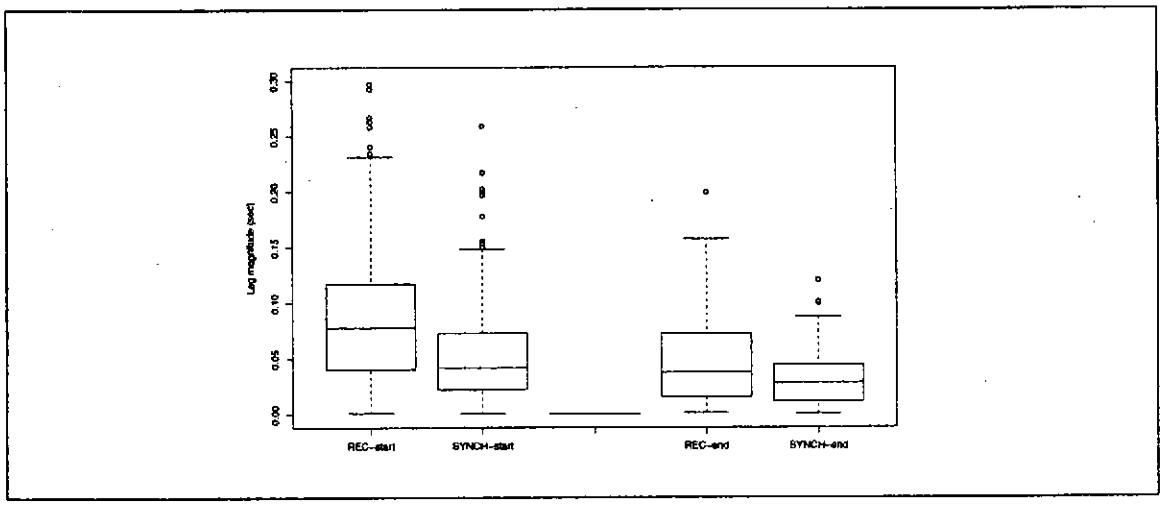

Figure 1: Absolute value of measured asynchrony at phrase start and ends.

Summarizing, subjects demonstrate clearly that they can achieve a high degree of synchrony. Median lag magnitude in the synchronous condition was 30 ms, (upper quartile 55 ms), but 56 ms (upper

quartile: 95 ms) in the recording condition. Remarkably, synchrony is maintained well across pauses when both speakers are "live".

3 EXPLOITING REDUCED VARIABILITY

We are currently examining the potential benefit of using Synchronous Speech as an elicitation technique in creating a speech database for concatenative synthesis. A large part of the art of concatenative synthesis consists of establishing a database of basic speech units from which elements are selected for concatenation. Where conventional approaches use diphones, demisyllables or phonemes as their basic concatenative units [1, 2], we take the whole word as our basic unit.

Word-based concatenative synthesis has not been actively pursued for many reasons: Contextual variation across words in different syntagmatic positions is, of course, very large indeed. Also, the number of words required for even basic TTS is very large compared with the number of distinct phonemes or diphones, which all but precludes having a database with multiple instances of each basic item, from which selections could be made in context-sensitive manner.

However our goals differ from those of many received approaches. Rather than shooting for optimal speech quality, we seek to optimize the procedure for constructing basic but functional TTS systems in languages for which little technical or linguistic support exists. Languages, that is, in which illiteracy is a substantial problem, and for which little inofrmation technology development has hitherto been done. In many cases, a word-based approach may be the simplest and fastest way of achieving reasonable quality synthesis: material will be gathered by having a reader read a large number of target texts (newspaper articles), and a semi-automatic alignment procedure will allow rapid extraction of word-sized units.

In pilot studies currently under way, we are assessing whether texts read in a synchronous condition produce material which can be used in a word-sized concatenation process. As Synchronous Speech promises to greatly reduce non-essential variability in speech, it may produce an advantage over unconstrained readings. Initial results of this assessment will be presented at the WISP workshop.

4 DISCUSSION

We have established that subjects can produce Synchronous Speech, and that by some measures at least, the inter- and intra-subject variability of the resulting speech is greatly reduced compared to a control condition. In the above study we focused on phrasing and pause placement. Ongoing studies are examining other variables in the temporal and intonational domains.

The utility of SS as an elicitation procedure in experimental phonetics remains to be tested. The study of intonation is one domain in which particular promise appears likely, as the single biggest problem in studying intonation has been the separation of the categorical or linguistic (expected to be preserved in SS) from the expressive or continuous (expected to be greatly reduced or shed).

Many open questions remain. We need to characterize the process of synchronization across speakers: Is there a two-way exchange of information, or is the process asymmetrical? How is information passed between co-speakers? Is visual information important? Synchronous Speech as a phenomenon is only just coming under scrutiny. Its potential as a tool remains, for now, unexplored.

References

[1] A. W. Black and N. Campbell. Optimizing selection of units from speech databases for concatenative synthesis. In *Proceedings of Eurospeech*, pages 581–584, 1995.

[2] Thierry Dutoit. *An Introduction to Text-to-Speech Synthesis*, volume 3 of *Text, Speech and Language Technology*. Kluwer Academic, 1997.

[3] Joseph Perkell and Dennis H. Klatt, editors. *Invariance and Variability in the Speech Processes*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[4] R. A. Rasch. Synchronization in performed ensemble music. *Acustica*, 43:121–131, 1979.