EXPLOITING PAUSES IN CONTINUOUS SPEECH RECOGNITION

Douglas O'Shaughnessy
INRS-Télécommunications, Université du Québec
3 Place du Commerce, Verdun, Québec H3E 1H6

1. Introduction

Most automatic speech recognition systems to date have required that speakers carefully pronounce their speech, in order to obtain good recognition performance. The task of automatic recognition of spontaneous, natural or conversational speech differs from that of careful or read speech in several ways, the most obvious difference concerning hesitation phenomena. In spontaneous speech, people often start talking and then think along the way. This causes spontaneous speech to have a variable speaking rate (both within and across sentential utterances); at times, such speech exhibits interruptions. The specific interruption phenomena studied in this paper are hesitation pauses (both filled and unfilled) and utterance restarts. Pauses are simple interruptions in the flow of speech, where a significant delay occurs in the delivery of the speech between words. In restarts, the speaker repeats or corrects some words (usually in addition to pausing).

A primary application of this study of hesitation phenomena lies in improving the performance of automatic recognizers, given an input of spontaneous speech (e.g., verbal conversations with computer databases). Speech researchers have often expressed interest in exploiting the intonation of spoken utterances in the recognition process, but have been deterred by the complex nature of how intonation (including pauses) relates to the text of an utterance. Even straightforward phenomena such as unfilled pauses (i.e., silence periods - which are generally easy to identify, if long enough) are not reliable indicators to the syntactic or semantic sentence structure of an utterance.

2. Speech database studied

In the context of our investigation into voice dialog access to databases, we are currently examining an application involving a simulated travel agent. A naive user (the speaker) is given the task of arranging a trip involving air travel via commerical airlines, by verbally interacting with a "computer travel agent." Thus, the user formulates verbal questions (and commands) on the fly, in a spontaneous fashion, as if in conversation with a travel agent. The current system does not reply verbally, but rather outputs information from a database onto a computer screen. The database is a version of the Official Airline Guide actually used by travel agents (it was furnished as part of a project supported by DARPA - the US Defence Advanced Research Projects Agency). The spoken data consists of more than 30 adult male and female speakers, each speaking about 30 utterances, each ranging in length from a few words to several dozen words. Many utterances are quite fluent, and exhibit no pause phenomena. However, about half contain pauses, and many have more than one hesitation phenomenon.

3. Previous studies on hesitation phenomena

In examining a corpus of speech produced by people spontaneously describing images, Levelt [1] found that 18% of the speech restarts occurred within a word, which was then corrected in the restart; i.e., the speaker paused in the middle of the "problem word" and restarted the utterance (e.g., "...go to the ye-, to the orange node"). In 51% of the cases, the speaker halted immediately after the word to be corrected, while 31% of the time the speaker stopped one or more words after the problem word (e.g., "...from green left to pink - er, from blue left to pink"). Most of the interruptions at word boundaries occurred at major syntactic boundaries. Within-word interruptions, on the other hand, did not even preserve syllable boundaries; i.e., speakers tended to stop immediately upon realizing that a problem existed, even if that meant stopping before a vowel could be pronounced in the current syllable. Levelt found that "uhh" occurred in 30% of restarts. He noted that uttering such a neutral sound (i.e., filling the pause) may help the speaker prevent an interruption by another speaker. The implication is that listeners often interpret unfilled pauses (i.e., silence) as a cue to start speaking, but they would not interrupt a filled pause. Levelt noted that restarts can be either marked prosodically by changes in intonation (between the speech before and after the pause) or unmarked prosodically (i.e., no change in intonation). Cases of simple mispronunciation tended to be unmarked, whereas lexical changes (replacement of a word with a different sense) were marked. While Levelt's work is of direct relevance here, it gives few quantitative details other than simple statistics of occurrence; in particular, F0 and durational distributions are rarely mentioned.

Deese [2] noted that hesitation pauses occur less often in planned (non-spontaneous) than unplanned speech. He defined such a pause as occurring in a syntactically inappropriate location and lasting between 100 ms and 300 ms. Our results below dispute these assertions: some occur at syntactic boundaries, and hesitation pauses can last well beyond 300 ms. Deese suggested that filled pauses lend an air of diffidence and humility, whereas unfilled pauses suggest assurance and superiority. In comparing planned and unplanned speech, he found that planned speech had more total pauses (10.3 per 100 words, vs. 8.8 for unplanned speech), while having fewer restarts (3.8 per 100 words, vs. 5.0 for unplanned speech). He found that pauses ranged from about 50 ms to 5 s (we found similar pause lengths here). Without giving quantitative results, he noted that long filled pauses tended to be segmented into syllables (i.e., "uhh umm uhh" rather than "uhhhhhh" or "ummmm"). Mispronunciations (words uttered incorrectly rather than chosen incorrectly) occurred at a rate of 1.5 per 10,000 words; mistaken words occurred at 2.5 per 10,000 words.

Hauptmann and Rudnicky [3] investigated ways in which humans differ in speaking to computers as opposed to speaking to other humans. They found that the average utterance to a computer was longer (6.1 words vs. 5.5 words to a human). Filled pauses occurred at a rate of 4 per thousand words when talking to a computer, and 15 per 1000 to a human. Almost all filled pauses occurred just before a definite reference to a name.

A few researchers have reported success in identifying major syntactic boundaries using prosodic means. 90% success rates are noted for English using prepausal lengthening [4,5] and for French using pitch patterns [6] and vowel durations [7]. Few details are described in this literature, however, and none exploit hesitation phenomena. Furthermore, the English results were based on read speech (not the spontaneous speech of our study) and assumed knowledge of the text corresponding to the speech (which would not be the case in an automatic recognition system).

A study of hesitations in spontaneous French speech [8] noted many similarities to English, and gives an idea how often hesitations occur. They found that a false start (as well as a simple word repetition) occurs on the average every 60 syllables, that a filled pause occurs on average every 22 syllables, and that an unfilled pause happens every 6.5 syllables on average. Thus, hesitation phenomena are very frequent in spontaneous speech and must be addressed in a recognition system attempting to handle such speech.

4. Experimental results

The grammaticality of a pause cannot be reliably separated based on silence duration. Both grammatical and ungrammatical sentence-internal pauses ranged from 100 ms to 3900 ms, with much overlap between the two classes. The 46 syntactic pauses examined averaged 900 ms (median = 800 ms), while the 22 ungrammatical ones averaged 715 ms (median = 600 ms). While there is a definite tendency toward longer silences at syntactic boundaries, a clearer distinction is found in the prepausal word; speakers tend to plan ahead for grammatical pauses and take action prior to the pause. One might expect that speakers would lengthen the final word before a planned pause (i.e., traditional prepausal lengthening) [9], while less often lengthening a word prior to a hesitation. Such a distinction did not occur in duration, but rather in the pitch contour. Very few ungrammatical pauses had a continuation rise in F0 (fundamental frequency) just prior to the pause, whereas 80% of the grammatical pauses were accompanied by a prior F0 rise of 10-40 Hz. These are reliable F0 patterns that can easily be extracted, and do not involve extracting F0 during unvoiced-voiced transitions (where F0 estimators have their greatest difficulties).

Not all pauses are as easy to locate as silences: filled pauses resemble words in continuous, spontaneous speech. A phonetic distinction is made here between filled pauses at major syntactic boundaries and those within syntactic units. Filled pauses at major boundaries were found in the range of 200-500 ms; those within syntactic units were shorter on average (e.g., 170-320 msec). Thus the ranges overlapped, but the syntactic nature of the filled pause could be distinguished by analyzing the silence periods adjacent to the filled pause: for the ungrammatical filled pause, a preceding unfilled pause was very brief (0-350 ms), as was any ensuing silence (0-500 ms). Each grammatical filled pause was preceded by a silence exceeding 275 ms; a long prior silence (> 700 ms) led to a relatively short filled pause (< 300 ms), whereas a short prior silence correlated with a long filled pause (> 300 ms). The spectral pattern of a filled pause was a uniform vowel during its duration (e.g., a steady schwa), possibly followed by the steady nasal /m/. Filled pauses all had falling (5-20 Hz) or flat F0 patterns, at relatively low F0 levels. Ones at syntactic boundaries tended to start higher in F0 and then fall, whereas filled pauses internal to a syntactic unit had lower F0 patterns. All had F0 ending in the bottom 15% of the speaker's F0 range.

5. Conclusion

Simple rules to exploit hesitation phenomena in recognition of spontaneous, continuous speech have been described. The acoustic measurements required (silence durations, F0 during vowels) are robust enough to be practical even for speech in noisy environments. The durations of pauses and F0 behavior in prepausal syllables allow reliable discrimination of whether a pause is occurring at a major syntactic boundary. Earlier results comparing pause statistics in English and French [8] suggest that the results here could be applied as well to French speech.

6. References

[1] W. Levelt, Speaking: From Intention to articulation. Cambridge, MA: MIT Press, 1989.

[2] J. Deese, Thought into speech: The Psychology of a language. Englewood Cliffs, NJ: Prentice-Hall, 1984.

[3] A. G. Hauptmann and A. I. Rudnicky, "Talking to computers: an empirical investigation," International Journal of Man-Machine Systems, vol. 28, pp. 583-604, 1988.

[4] C. W. Wightman and M. Ostendorf, "Automatic recognition of prosodic phrases," in Conference Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 321-324, 1991.

[5] P. J. Price, C. W. Wightman, M. Ostendorf, and J. Bear, "The use of relative duration in syntactic disambiguation," in Conference Proceedings of the International Conference on Spoken Language, pp. 13-16, 1990.

[6] J. Vassière, "On automatic extraction of prosodic information for automatic speech recognition system," Conference Proceedings of Eurospeech-89, pp. 202-205, 1989.

[7] N. Carbonell, "On the use of prosodic knowledge for continuous speech recognition and understanding," Conference Proceedings of Eurospeech-89, pp. 522-252, 1989.

[8] F. Grosjean et A. Deschamps, "Analyse des variables temporelles du français spontané," Phonetica, vol. 28, pp. 191-226, 1973.

[9] T. Crystal and A. House, "Segmental durations in connected-speech signals," Journal of the Acoustical Society of America, vol. 84, no. 4, pp. 1553-1585, 1988.