RECOGNIZING DISFLUENCIES IN SPONTANEOUS SPEECH

Douglas O'Shaughnessy
INRS-Telecommunications, 16 Place du Commerce
Nuns Island, Quebec, Canada H3E 1H6

1. INTRODUCTION

Most previous acoustic analysis of speech has examined data from speakers who carefully pronounce their speech, usually by reading prepared texts. Natural spontaneous 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 interruptions; the specific interruption phenomena studied in this paper are pauses and restarts. Pauses can be either unfilled (silence) or filled with a speech sound (usually "uhh" or "umm" in North American English). Restarts (or false starts) are interruptions in the flow of speech, where the speaker (usually after a brief pause) reiterates a portion of the speech immediately preceding, with or without a change. The repetition can range from a portion of a syllable up to several words. In the case of a change, the modification may be either a substitution of a new word (in the place of a fully- or partially-spoken previous word) or an insertion of a word in a word sequence (with the sequence containing the new word being uttered again).

This paper concerns the acoustic analysis of pauses and restarts in spontaneous speech, from the point of view of their automatic location via acoustical analysis. A large database of spontaneous speech was analyzed in terms of duration and fundamental frequency measurements, as well as spectral analysis. For recognition purposes, a simple spectral analyzer was used to identify repeated words.

The pauses and restarts are described acoustically, with a view toward automatic recognition, to ensure their proper elimination from consideration in speech recognition systems. A primary application of this study lies in improving the performance of automatic speech recognizers, for applications that must accept an input of spontaneous speech (e.g., verbal conversations with computer databases). For such purposes, we wish to eliminate filled pauses and one version of any repeated words (or parts of words), and in the case of changed words, we wish to suppress the original unwanted words, so that the recognizer will operate on only a sequence of desired words. Thus, we examine here the relationship of pauses and restarts to intonation, and do so in a fashion that should allow direct exploitation in automatic recognition systems accepting spontaneous, continuous speech.

Speech researchers have often expressed interest in exploiting the intonation of spoken utterances in automatic recognition algorithms, but have been deterred by the complex nature of how intonation relates to the text of an utterance. Various aspects of the intonation employed in a restart allow it to be identified as a restart, and furthermore allow suppression of the undesired words in many cases. Pauses are more simple to locate, but unfilled pauses can be confused with phonemic stop closures, and filled pauses can be confused with monosyllabic words.

Within-utterance hesitations can cause significant difficulties for automatic speech recognizers, which usually make no provision for pauses at random locations or for repeated words or parts of words. Automatically determining which words (or parts of words) are being replaced in a speech repair and locating filled pauses could help automatic recognizers avoid textual errors in the output. In virtually all current recognition systems, words repeated in a false start are either simply fed as word hypotheses to the textual component of the recognizer or cause difficulties in having a proper interpretation in the language-model component (since the language model is invariably trained only on fluent text). Similarly filled pauses appear as actual words in the textual output.

2. PREVIOUS STUDIES

Acoustical analyses of disfluencies with a view toward speech recognizers are rare [2]. Previous work on restarts has dwelled almost exclusively on the length of the word-repeat sequences (and occasionally on the pause duration). Most of the work on restarts that has been reported in the literature has treated the phenomena in a general qualitative or overly simple quantitative fashion. As far as we know, no reports have previously linked the intonational cues of both F0 (fundamental frequency) and duration to restarts in a way that could be useful to automatic speech recognition. Indeed, very few recognition systems use intonational cues, especially F0, at all. In this paper, we examine how these latter parameters could be exploited directly.

Recently an attempt was made to automatically detect and correct restarts in spontaneous speech [2]. Looking at an enlarged version of our own database, the authors examined 10 000 utterances, of which 607 were found to have restarts. In utterances longer than nine words, a significantly high 10% had restarts. 59% of the restarts involved only one word (whose deletion would render the sentence fluent); 24% involved two words (or word fragments); 8% involved three words, etc. Of the one-word restarts, the majority (61%) involved a word fragment, 16% involved the repetition of a word, 7% involved inserted words, and 9% concerned replacement words. The majority of the two-word restarts were either a straight repeat of two words or a replacement of the second word, while 19% involved inserted words, and 10% involved a replacement of the first of the two words.

3. SPEECH DATABASE

In this paper, we examine disfluencies in a standard speech database (used by several speech recognition research groups in North America), ranging from pauses to simple restarts (involving only the repetition of 1–2 words) to complex restarts (where, instead of simply repeating words, one substitutes a new word for an unwanted one).

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 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 spoken data consists of 42 adult male and female speakers, each speaking about 30 different utterances, each ranging in length from a few words to several dozen words (median length of about 12 words).

In the approximately 1000 utterances examined (from many different speakers, each containing an average of about thirteen words), there were 60 occasions where the speaker simply repeated words or portions of words, 30 cases of inserted words, and 25 occurrences of new words substituted for prior spoken words (or word parts). Thus, approximately

10% of the utterances (a percentage consistent with the parallel study of [2]) had a restart.

4. ANALYSIS METHOD

Hardcopy displays were made of all utterances containing restarts (as determined by listening and transcribing each utterance), in sections of 3-5 seconds at a time. Each display contained a waveform (amplitude vs. time) and a narrowband spectrogram (showing 0-2 kHz). Time resolution in these displays ranged from 44 to 78 mm/s; the frequency axis showed 39 mm/kHz. These displays were manually segmented into words and syllables, and F0 contours were obtained by tracing strong harmonics in the middle of the first or second formant.

5. ACOUSTICAL ANALYSIS RESULTS

Actual unfilled pauses (as distinct from silence in stop closures) were as short as 40 ms and as long as several seconds. The 149 unfilled pauses examined at syntactic boundaries averaged 760 ms (median = 490 ms), whereas the 92 pauses within major syntactic units averaged 490 ms (median = 270 ms).

Filled pauses resemble short words in continuous, spontaneous speech. Filled pauses at major syntactic boundaries had durations in the range of 200-500 ms; those within syntactic units were shorter on average (170-320 ms). 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 (if any) was very brief (less than 350 ms), as was any ensuing silence (less than 500 ms). Each grammatical filled pause was preceded by a silence exceeding 275 ms; a long prior silence (more than 700 ms) led to a relatively short filled pause (less than 300 ms), whereas a short prior silence correlated with a long filled pause (more than 300 ms).

The spectral pattern of a filled pause was a uniform vowel during its duration (e.g., a steady schwa), sometimes 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.

As for false starts, when a word was simply repeated (as is) in a restart, it had virtually the same prosodics (i.e., same duration and pitch) in both its instances in most cases, but there were a number of times where the repeated word had less stress (i.e., shorter duration and lower pitch). When a word was changed (i.e., a substitution or insertion) in the restart, on the other hand, its second instance was virtually always more stressed (i.e., longer duration and higher pitch).

In the case of restarts where the speaker stopped in the middle of a word and simply "backed up" and resumed speaking with no changed or inserted words, the pause lasted 100–400 ms in 85% of the examples (with most of the remaining examples having a pause of about 1 second in duration). About three-fourths of the interrupted words did not have a completion of the vowel in the intended word's first syllable (e.g., the speaker usually stopped after uttering the first consonant). In virtually all examples, the speaker completed at least 100 ms of the word, however, before pausing for at least 100 ms. When the pause occurred at a word boundary, the words repeated after the pause were characterized by two situations: either a straight repetition with little prosodic change (this happened especially when a lengthy pause intervened), or a repetition where the repeated words shortened up to 50%.

In the case of a word being substituted or inserted into the word sequence in the restart, the substituted/inserted word received a large stress (relatively long duration and rise in F0) in examples where the new word added significant semantic information, but did not in examples where the new word was redundant in terms of the prior context (e.g., if the new word was a synonym of an immediately previous word). As for the repeated words (after the pause) prior to the inserted word, function words showed little or no shortening, but usually had lower F0; on the other hand, content words here exhibited significant shortening and lower F0 (the shortening here was about 50% for short words less than 300 ms, and about 100-200 ms for longer words). Such prosodic change only applied to non-prepausal words, because words immediately prior to a pause were often subject to significant prepausal lengthening.

6. RECOGNIZING RESTARTS

Since pauses involved in restarts were generally shorter than other pauses [1], we could suggest a simple rule of "pause < 400 ms -> restart." For our database, such a rule will correctly identify 70% of restarts, but will give 35% false alarms (i.e., incorrectly claiming as restarts those grammatical pauses which are shorter than 400 ms). While this performance is well above chance, it is clear that pause duration alone is not a reliable cue to a simple restart. Also, restart pauses at the very start of utterances were quite variable in duration (the 400 ms rule is more reliable when applied to pauses found after 3 syllables of an utterance). Obviously, the spectral-time detail on either side of a pause must be examined to verify whether a restart is present.

Since most restarts are simple repetitions, looking for identical spectral-time patterns (of up to 3 syllables in length) on either side of a short pause will greatly increase the restart recognition accuracy. For simple repetitions, the scope of spectral analysis is very limited: one need only look at about 2–3 syllables before and after each candidate pause. Very few simple repetitions repeated more than three syllables (a significant portion of complex restarts, on the other hand, involve more than three syllables). If a close spectral match is found and the pause exceeds a low threshold (e.g., 120 ms – to avoid confusion with stop closures), we declare that the pause is a simple restart, and that one version (usually the first) of the matching syllables should be excluded from consideration in any ensuing recognition process. We were very successful in automatically recognizing such simple restarts.

Recognizing restarts with changed words appears to be much more difficult than identifying simple restarts. We look for a short pause (again < 400 ms), followed by a spectral-time pattern containing 1–2 syllables corresponding to a portion of the speech immediately prior to the pause. However, there are many possibilities here and many of them have spectral and prosodic patterns that resemble fluent speech (i.e., speech without repeated or substituted words, but having pauses). For example, after the pause in such a restart, the immediately ensuing word(s) may be the added/substituted ones, or there may be one or two repeated words (from before the pause). The added/substituted words may be as short as one or as long as six syllables.

ACKNOWLEDGMENTS

This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada.

REFERENCES

[1] O'Shaughnessy, D.: "Recognition of hesitations in spontaneous speech," ICASSP-92, pp. 593–596, March 1992.

[2] Bear, J.; Dowling, J.; Shriberg, E.: "Integrating Multiple Knowledge Sources for Detection and Correction of Repairs in Human-Computer Dialog." Proc. Assoc. Computational Lingusitics, pp. 56–63, 1992.

[3] Shriberg, E., "Intonation of clause-internal filled pauses," Proc. ICSLP-92, pp. 991–994, Oct. 1992.