HANDLING FALSE STARTS IN RECOGNITION OF 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 restarts (or false starts), which 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 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 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 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 restarts to intonation, and do so in a fashion that should allow direct exploitation in automatic recognition systems accepting spontaneous, continuous speech.

Within-utterance hesitations can cause significant difficulties for automatic speech recognizers, which usually make no provision for repeated words or parts of words. Automatically determining which words (or parts of words) are being replaced in a speech repair 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).

2. PREVIOUS STUDIES

Acoustical analyses of disfluencies with a view toward speech recognizers are extremely rare. (To our knowledge, the only such work is recent and found in [1–2].) 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 [3–5]. 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 [1–2]. 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).

These authors [1–2] tried to automatically locate and correct these restarts, first using text alone (assuming that a speech recognizer could provide a correct transcription) and then using cues from the speech itself. Based on simple pattern matching of the text alone (e.g., looking for repeated words, cue words, and simple syntactic anomalies), their algorithm had a relatively high error rate for location: missing 23% of the utterances that had restarts and producing false alarms in 38% of the proposed cases. The rate for correcting the restarts (for the properly located ones) was only 57%. After inclusion of a language model, they were able to detect 85% of the restarts, based on text input. They noted that simple repeats usually have a significant pause (mean = 380 ms) between the repeated words, whereas actual (intended) repeated words (e.g., 'flight five one one') have very brief (if any) pause [6a]. They further noted that, for restarts where one word was repeated with a new inserted word prior to it (e.g., '... flight [pause] earliest flight...'): 1) if a pause was present adjacent to the new, inserted word, it was found before the word, and 2) the new word had a higher peak F0 than the preceding word.

The approach of these last authors is similar to ours (and we use the same database), but we address the question of automatic location of restarts, and without prior knowledge of neither the word boundaries nor the identity of the words. In practical recognition applications, one certainly cannot assume such prior knowledge. Thus it is quite practical to examine how such disfluencies can be found directly from an acoustical analysis of the audio (speech) signal.

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 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 [4–5]) 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

With 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.

We concentrate now on simple repeat restarts (i.e., those where words or parts of words were simple repeated, with no change in their content), because they were the most promising to recognize automatically. These can be divided into utterance-initial restarts and utterance-medial ones. At the very start of an utterance, speakers often start saying something and immediately stop, having uttered only a few syllables or even a fraction of a syllable (e.g., 'Wh- @ what I want...,' 'I'd @ I'd like to know...', where '@' represents a pause). The pauses in these cases were very variable, ranging from 80 ms to a few seconds, unlike the vast majority of simple repeat pauses in utterance-medial position, which were in the range 100–400 ms. When the pause at the start of an utterance exceeded 400 ms, after speech of fewer than three syllables, the speaker usually restarted his utterance rather than continuing as if no pause intervened.

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 syllable or as long as six syllables. Due to the difficulty of distinguishing complex restarts from fluent pauses, a simple algorithm for identifying such restarts awaits further research.

7. CONCLUSION

This paper has detailed the extent of prosodic phenomena in speech restarts in a multi-speaker database of spontaneous, continuous speech, and has given intuitive explanations for them, based on a theory of using prosodics to cue semantic information to a listener. Based on the acoustic data, ways have been described as to how to attempt to recognize these phenomena in the context of an automatic speech recognizer.

Simple restarts can be distinguished acoustically, via an analysis of duration, F0 and spectral detail in the neighborhood of a pause. Restarts with changed words may be distinguishable, but the required analysis will need to be much more complex. It will require a detailed examination of the pitch and durations of the pauses and adjacent words, along with acoustic recognition of words or syllables. Unfortunately, the wide variety of possibilities seen in this study for restarts with a modification does not suggest a simple recognition algorithm at this time.

ACKNOWLEDGMENTS

This work was supported in part by grants from the Natural Sciences and Engineering Research Council of Canada, the Fonds pour la Formation de Chercheurs et l'Aide a la Recherche (Quebec), and the Canadian Networks of Centres of Excellence program (IRIS).

REFERENCES

[1] Shriberg, E.; Bear, J.; Dowling, J.: "Automatic Detection and Correction of Repairs in Human-Computer Dialog." DARPA Speech and Natural Language Workshop. Arden House, N.Y., 6 pages, Feb. 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, June 1992.

[3] Hieke, A.: "A Content-processing view of hesitation phenomena." Language and Speech, vol. 24, Part 2, pp. 147–160, 1981.

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

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