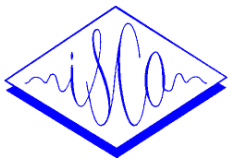

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

ANALYSIS OF FALSE STARTS IN SPONTANEOUS SPEECH

Douglas O'Shaughnessy

INRS-Télécommunications
Université du Québec
16 Place du Commerce, Verdun, Quebec H3E 1H6 Canada

ABSTRACT

A primary difference between spontaneous speech and read speech concerns the use of false starts, where a speaker interrupts the flow of speech to restart his utterance. The acoustic aspects of such restarts in a widely-used speech database were examined here. Identifying the type of restart in such cases could improve the performance of an automatic speech recognizer, by eliminating from consideration some hypotheses based on spectral analysis. Results are given here which could allow simple identification of most restarts and their type.

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 a variable speaking rate (both within and across sentential utterances), and such speech often exhibits interruptions. The specific interruption phenomena studied in this paper are restarts, 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 in the words. The repetition can range from a portion of a syllable 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).

Thus, this paper concerns the acoustic analysis of restarts in spontaneous speech. A large database of spontaneous speech was analyzed in terms of duration and fundamental frequency measurements. 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.

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.

Within-utterance hesitations can cause significant difficulties for automatic speech recognizers, which usually make no provision for unpredicted pauses or 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 output textual errors. 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).

Our previous work reported on hesitation phenomena in spontaneous speech in general, and focussed on pauses (both filled and unfilled). The current paper concentrates on the more difficult task of recognizing restarts, rather than simply identifying the syntactic nature of pauses. We present here a more comprehensive analysis of restarts than found in the current literature, including examination of the duration and pitch of the words surrounding the pause in a restart. Such an analysis yields very useful information for speech recognition. In addition, we give intuitive explanations for the phenomena, based on a theory of using prosodics to cue semantic information to a listener.

2. PREVIOUS STUDIES ON RESTARTS

Acoustical analyses of restarts with a view toward speech recognizers are extremely rare (or nonexistent). (To our knowledge, such work has not been presented in a journal or at a conference dealing with speech recognition, with the exception of a very recent workshop presentation [6].) 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. For example, the linguistics literature describes where such restarts are likely to be found in broad terms of syntax and semantics, but this literature gives little quantitative detail. The cognitive psychology literature gives simple statistics regarding restarts, in terms of frequency of occurrence. 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.

Heike describes restarts as serving one of two functions: stalling and repairing [7]. In his view, speakers try to optimize speech communication efficiency by compromising between fluency and speed. Since thought is necessary to organize one's speech, it appears that each speaker has a choice between speaking slowly enough (with pauses) to permit enough thought to avoid making any mistakes that might require correction, on the one hand, and minimizing pauses (for efficiency and speed), on the other hand, while risking restarts. (In our own database, it is clear that some speakers adopt the first option, speaking slowly with many pauses but few restarts, whereas others have few pauses, but a greater number of restarts.)

Repeated words in restarts, according to Heike [7], can be either retrospective (correcting) or prospective (anticipatory). The latter are similar to filled pauses and are used to gain time while thinking, while the former serve to recall words (by repeating them) that occurred too long ago in the speech. As an example of a retrospective repeat, when one says "I want the (long pause) the table," one usually repeats the word "the" after the pause, because it is very closely related to "table" in the syntactic structure of the sentence (e.g., they form a noun phrase). In general, stalling can be accomplished by pauses, prospective repeats, or even prolonging syllables, whereas repairs are done by corrective restarts or retrospective repeats. Repairs include corrections due to mispronunciation, due to wanting changes in words (substitution or addition of words) or syntax, or due to inappropriate intonation.

Examining one-minute spontaneous speech samples, Heike found that 90% of hesitations were stalls with only 7% repairs (and 3% parenthetical remarks). Among the repairs, 19% involved substitution of new words, 18% involved addition of new words, 37% concerned changes in syntactic structure, 15% involved mispronunciations, and 11% related to intonation changes.

In examining a corpus of speech produced by people spontaneously describing colored images, Levelt [3] found that 18% of the restarts occurred within a word, which was then corrected in the restart; i.e., the speaker paused in the middle of the incorrect 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 incorrect word (e.g., "...from green left to pink - er, from blue left to pink"). Most of the interruptions that occurred at word boundaries coincided with 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 there was a problem, even if that meant stopping before a vowel was pronounced in the current syllable. Levelt found that the filled pause "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 tend not to interrupt a filled pause. Levelt noted that restarts can be either marked prosodically by changes in intonation (in 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 interest here, it gave few quantitative details other than simple statistics of occurrence; in particular, F0 and durational distributions were rarely mentioned.

In comparing planned and unplanned speech, Deese [4] noted that planned speech had fewer restarts (3.8 per 100 words, vs. 5.0 for unplanned speech). Much rarer than pauses, mispronunciations (intended words uttered incorrectly, rather than chosen incorrectly but properly pronounced) occurred at a rate of 1.5 per 10,000 words; mistaken words occurred at 2.5 per 10,000 words.

A study of hesitations in spontaneous French speech [5] noted many similarities to English. It was found on the average that a false start (and also a simple word repetition) occurred every 60 syllables. Thus, hesitation phenomena can be very frequent in spontaneous speech and should be addressed in a recognition system attempting to handle such speech.

A very recent report in the literature describes an attempt to automatically detect and correct restarts in spontaneous speech [6]. Looking at an enlarged version of our own database, they 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 major-

ity 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.

Shriberg et al [6] 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.

3. SPEECH DATABASE

In this paper, we examine false starts 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 [6]) 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

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 are generally shorter than other pauses, we could suggest a simple rule of "pause $<$ 400 ms $\rightarrow$ restart." For our database, such a rule will correctly identify 70% of restarts, while giving 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. 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. If a close spectral match is found and the pause exceeds a low threshold (e.g., 80 ms – to avoid confusion with stop closures), we would 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.

Recognizing restarts with changed words appears to be much more difficult than identifying simple restarts. We suggest looking 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 clear and simple algorithm for identifying such restarts awaits further research.

7. CONCLUSION

In the paper, I have detailed the extent of prosodic phenomena in speech restarts in a multi-speaker database of spontaneous, continuous speech, and have given intuitive explanations for them, based on a theory of using prosodics to cue semantic information to a listener. I have also given specific suggestions (based on the acoustic data) as to how to attempt to recognize these phenomena in the context of an automatic speech recognizer.

It was shown that simple restarts (i.e., those without inserted or substituted words) could be distinguished acoustically, via an analysis of duration, F0 and spectral detail in the neighborhood of a pause. We expect to be able to automatically identify such restarts with an accuracy exceeding 80%, while keeping false alarms to below 10%.

Restarts with changed words may be distinguishable, but the required analysis will need to be much more complex, and beyond the immediate scope of this paper. 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, from the Fonds pour la Formation de Chercheurs et l'Aide a la Recherche (Quebec), and from the Canadian Networks of Centres of Excellence program (Institute for Robotics and Intelligent Systems).

REFERENCES

[1] J. Vassiere. "On automatic extraction of prosodic information for automatic speech recognition system." Eurospeech-89, vol. 1, pp. 202–205, 1989.

[2] W. Ward. "Understanding spontaneous speech: The Phoenix system." IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). pp. 365–368, 1991.

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

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

[5] F. Grosjean and A. Deschamps. "Analyse des variables temporelles du francais spontane." Phonetica. vol. 28, pp. 191–226, 1973.

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

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