THE PROSODY OF SPEECH ERROR CORRECTIONS REVISITED

Stefanie Shattuck-Hufnagel and Anne Cutler

Speech Group, Research Laboratory of Electronics, MIT, Cambridge, MA
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands)

ABSTRACT

A corpus of digitized speech errors is used to compare the prosody of correction patterns for word-level vs. sound-level errors. Results for both peak F0 and perceived prosodic markedness confirm that speakers are more likely to mark corrections of word-level errors than corrections of sound-level errors, and that errors ambiguous between word-level and sound-level (such as boat for moat) show correction patterns like those for sound level errors. This finding increases the plausibility of the claim that word-sound-ambiguous errors arise at the same level of processing as sound errors that do not form words.

1. INTRODUCTION

Errors in spoken utterances are systematic and highly constrained, and hence have long been called in evidence to support models of the cognitive process of planning speech for production (see e.g. [1,2,3,4,5,6,7]). Although the study of language production has in recent years seen a welcome growth of experimental approaches (see e.g. [8,9]), error phenomena have not thereby lost their value, and indeed the conclusions based on laboratory experiments and on analysis of natural error corpora have tended to converge. Most corpora of spoken errors used for this purpose have been collected by listeners writing down what they hear, and this method has certainly proved fruitful for analysis of, for example, the positional constraints on which elements in an utterance can interact, the types of elements which move or change in errors, and the ordering of processing steps, as for example in the accommodation of the phonetic shape of morphemes to elements introduced by errors ("busses back" for "backs busses"). However, as many investigators have pointed out, recording errors by writing them down is not a foolproof method of capturing all of the information about error patterns that one would like to have. It may not accurately sample the distribution of error types (if some types are more detectable, or easier to remember, than others), it does not capture potentially important sub-phonetic variation in the acoustic shape of the utterances [10], and it does not reflect an aspect of error utterances which is potentially of critical importance: their prosodic structure. These kinds of information are only available for errors which are acoustically recorded and can be digitized, heard repeatedly and analysed instrumentally. Since such information has the potential to distinguish between competing models of speech production planning, the development of acoustically-recorded error corpora is an important goal.

Several recorded corpora exist ([11], [12]), but these collections are often small, because of methodological difficulties. For example, harvesting appreciable numbers of errors from existing speech databases requires listening to prohibitively large quantities of utterances; although it is not unusual to hear a number of errors in the course of a day of listening to speech, their rate of occurrence per utterance is relatively low [13]. Disfluencies such as unexpected pauses, unintelligible fragments, repetitions and restarts are not hard to find, but errors in the sense of misorderings, omissions and additions of constituents are rarer. Understanding the constraints on this set of events is, again, of particular interest for theoretical reasons, and the variety of errors of this type that are observed to occur is quite wide. Thus, a large corpus is needed to permit accurate analysis of their distribution patterns, but large numbers of errors have not been collected. In addition, existing recorded corpora are often not generally available to the research community, in part because they were collected before the widespread availability of digitization technology, and in part because of the lack of an established distribution mechanism.

To remedy this situation, we have embarked on the development of a digitized speech error corpus, using a number of sources of recorded speech. Currently the MIT Digitized Speech Error Corpus (MIT-DSEC) contains more than 500 errors, and is growing daily. It relies principally on the harvesting of errors from speech that is being monitored carefully for other reasons, eliminating the need to devote long hours of listening time which is not productive of anything but the detection of an occasional error. This corpus will eventually be publicly available as digitized speech files (and the first author welcomes further contributions from the research community; contact stef@speech.mit.edu).

One of the most important possibilities opened up by this expanding corpus is the analysis of the prosody of speech errors and of the speech surrounding them, including detection, interruption and correction patterns. Some errors go undetected by the speaker; many others are detected. A detected error is not always corrected, but again, many are. When a speaker decides to correct an error, there are again two options - call the correction to the attention of listeners, or not. This latter dichotomy in the way speakers correct errors was first pointed out by Goffman [14]; prosodic analyses of tape-recorded errors by Cutler [12] and Levelt and Cutler [15] established that it was well reflected in the acoustic record. Cutler's [12] initial test of this suggestion was carried out on a relatively small subset of the initial corpus of taped errors from spontaneous speech referred to above. This analysis compared F0 and amplitude contours of error material and the corresponding correction material, and found a bimodal distribution: error and correction either differed substantially on these measures, or hardly differed at all. The difference measures did not distribute evenly along a continuum,

but fell into two apparent categories, as Goffmann on the basis of informal listening had proposed. Cutler referred to these two categories as "marked" corrections (the speaker calls the listener's attention to the correction by the use of very different prosody) and "unmarked" corrections (attention is not called to the correction). Cutler also observed that the prosodic marking of correction patterns was different for corrections of word-level errors than for corrections of sound-level errors. Word-level errors (e.g. "substitution" for "advertising"), involve mis-selection of an item from the speaker's stored lexicon, or a misordering of word-level elements. In contrast, sound-level errors (e.g."shong-lort" for "long-short") involve mis-selection or misordering of individual sound segments or strings of segments smaller than the morpheme. Nooteboom [3] had reported that the correction patterns for word-level errors were different from sound-level errors in a corpus of Dutch errors collected in written form; Cutler [12] extended this observation with the finding that, in the small set she analysed, all of the tokens judged perceptually to be prosodically marked fell into the word-level set, while unmarked corrections occurred with both word-level and sound-level errors.

Cutler argued that speakers are more concerned to mark corrections of errors which might mislead listeners. Follow-up analyses of a corpus of 299 taped Dutch errors by Levelt and Cutler [15] supported this claim, by showing that corrections of errors are more likely to be marked than repairs which simply substitute an alternative formulation, and errors involving substitution of an opposite (e.g. "left" for "right") are more likely to be followed by a marked correction than errors involving mis-selection from a larger set (e.g. "green" for "yellow"). The difference between word- and sound-level corrections thus arises because saying the wrong word is more likely to mislead than saying the wrong sound.

One interesting aspect of the distinction between word-level and sound-level errors is the fact that many errors are ambiguous on this dimension. That is, for some errors - e.g. "keep Tar Talk on the air" for "keep Car Talk on the air" - it is not entirely clear whether the error involved substitution of a word (here, "Tar" for the target word "Car"), or substitution of a phonetic segment (here, /t/ for the target consonant /k/). Some authors have claimed [4,5] on the basis of such word/sound-ambiguous errors that there is feedback from phonetic to lexical processing in speech production, despite the considerable evidence to the contrary from experimental studies [9]. However, it may obviously be the case that such ambiguous errors constitute a heterogeneous class: some are word-level errors, some are sound-level errors. If so, then the prosodic characteristics of such errors may offer an opportunity to refine the classification. The present study attempts an initial step towards this refinement by asking whether correction patterns within this class of ambiguous word-/sound-level errors also show evidence of a bimodal distribution of marking. For a subset of the errors in the MIT-DSEC, we compare F0 peak values as well as perceived markedness for error and correction.

2. METHOD

2.1 Corpus

The MIT-DSEC: The corpus includes errors from a number of sources, including a) existing recorded corpora, e.g. that of Cutler [12], b) errors recorded in the broadcast studio or via line-in from radio broadcasts, e.g. the BU FM Radio News corpus, c) errors from large speech corpora developed for automatic speech recognition efforts, e.g. the ATIS corpus of information queries from air travel professionals, and d) errors that occur in the course of laboratory speech elicitation experiments carried out for other purposes. In the future we plan to add a subsection of e) errors that occur during error elicitation experiments using specially-designed stimuli with tongue-twister-like characteristics. Each of these sources has its advantages and disadvantages. The Cutler corpus is drawn from truly spontaneous speech, since it was recorded from ongoing conversations using a mini-recorder. This overwhelming advantage comes at a price: the speech signal is often overlapped with background noise or other speech, and when recorded from radio broadcasts by microphone is sometimes hard to understand. The BU FM Radio News speech corpus and other on-line recorded radio speech have the great advantage of a high quality signal, but again the target speech signals sometimes also overlap with background babble from other speakers or music, and it is not always clear (e.g. during fundraising broadcasts) whether a speaker is reading or speaking spontaneously. The ATIS corpus provides a high-quality signal, and the speech is fully spontaneous, but the speakers use a very limited vocabulary, made up mostly of words like "flight" ,"from", "to", city names, days of the week, months and numbers. Finally, errors from elicitation experiments, whether they use normal sentences or specially-designed tongue-twisters, may not invoke all of the processing that is used in normal conversational speaking. However, all of these errors (and their corrections) illustrate part of the behavioral repertoire of speakers; it is hoped that, together, they will reflect the distribution of error patterns across a variety of speaking situations, speakers and tasks.

Errors from the Cutler corpus, the FM radio corpora and the lab speech corpora were received in audio tape form, and were digitized at 10K using the Klattools. They were subsequently reformatted for analysis using the Xwaves signal analysis software from Entropic Inc., to obtain aligned displays of the spectrogram, waveform display and F0 estimates for each utterance. Errors from the ATIS corpus were received in Xwaves-compatible format, and required no reformatting.

The Correction Corpus: The subset of MIT-DSEC errors described in this paper includes those which satisfy a number of specific criteria. First, the signal quality must be high, so that F0 estimates are likely to be reliable. However it should be noted that F0 trackers are imperfect tools, and we found a number of cases of pitch halving or even failure to track an F0 for an entire syllable. Second, the speaker must have produced enough of the vowel in the error word that the F0 tracker can produce an estimate for that region of the wave form. Third, the speaker must have produced a correction that provided an appropriate word/syllable to measure; this again was not always the case, as when the correction reflected an entire change of plans with a new phrase, new sequence of words and new prosody of its own. The final criterion was that the nature of the error was clear. The Correction Corpus currently numbers 90 tokens.

2.2 Measurements/Labelling

F0 measures: The F0 display for the error-correction region was displayed on the work-station screen (at a constant scale for all utterances), and the F0 peak for the error element and the corresponding element in the correction was determined by hand. Challenges included a) how to deal with the distortions introduced by voiceless consonants adjacent to the target vowel, which could result in spuriously high or low F0 values, b) what to do about errors in which the interruption truncated the utterance in the middle of the relevant vowel, ensuring that it did not reach its target F0 level, and c) how to handle multiple successive errors, like "bih---dee---being done", in which a comparison between the second error and the final correction is actually a comparison between one 'correction' attempt (which failed) and another (which succeeded). The effect of adjacent consonants was most problematical when e.g. an initial /s/ or /t/ resulted in a rapidly-falling F0 region at the onset of the vowel, and a potentially spuriously high F0 reading. This problem was resolved by taking the peak F0 value within the region in which the rate of F0 change was reasonable; since all F0 tracks were displayed at the same scale, this threshold was constant across tokens. The problem of truncated vowels was not resolved; the observed F0 peak was taken as an estimate of the intended F0 peak, and this probably resulted in a number of tokens being labelled with an F0 increase in the correction which might not have showed an increase if the error word or syllable had been completed. However, there was no obvious difference in the number of truncated vowels across the word-error, sound-error and ambiguous-error tokens. Finally, for multiple successive errors the final correction and the immediately-preceding error were compared; earlier errors were ignored.

Perceptual measures: A perceptual estimate of the markedness of each correction was made, to investigate the possibility that patterns in raw F0 differences might be usefully supplemented by an understanding of whether those differences produce an impression of marking for the listener. The listener had three responses available: marked, unmarked and questionable.

Phonological labelling: To understand more fully the results of the F0 and perceived markedness analyses, it is useful to have a transcription of the intonational phonology of both the original error and the correction, i.e. in the intonational framework developed by Pierrehumbert [16] and Beckman and Pierrehumbert [17], the pitch accents, phrase tones and boundary tones that define the intonational phrasing of the utterance. The ToBI transcription system [18,19] provides a tool for hand-labelling this information. The MIT-DSEC is currently being transcribed using the ToBI system; results will be reported in future publications.

3. RESULTS

Results for both the F0 and perceived markedness measures were, first, consistent with the prediction from Cutler [12,15], that the correction of a word-level error is likely to be produced with a more extreme F0 value than the error, and that this difference is perceived as a kind of prosodic markedness, while sound-level errors are less likely to be marked in this way. Second, errors which are ambiguous between word and sound errors (because the target and the error are both words, and the target word and error word differ only by one sound, as in "Tar---Car Talk") show correction patterns similar to those for sound-level errors.

Peak F0 results: The difference between the peak F0 values for an error and its correction was greater than 15 Hz significantly more often for word-level errors than for sound-level errors (chi-square value of 4.13 is consistent with p less than or equal to .05). In addition, the pattern for word-sound-ambiguous errors was indistinguishable from that for sound-level errors. Although the difference between the word-level corrections and the word-sound ambiguous corrections was only marginally significant ($p < .1$) here, the overall pattern of the results is consistent with the view that sound-error corrections are different from word-error corrections, and that ambiguous-error corrections behave like sound-error corrections with respect to F0 differences.

Perceived markedness results: A similar pattern was observed for judgments of perceived markedness. Although corrections of word-level errors were not significantly more likely to be judged as marked than corrections of sound-level errors ($p > .1$), ambiguous-error corrections were significantly different from word-error corrections ($p < .05$) and could not be distinguished from sound-level corrections ($p > .1$) on this dimension.

4. DISCUSSION

The results of this preliminary analysis of 90 errors from the developing MIT-DSEC support the proposal by Cutler [12] that speakers are more likely to mark corrections of word errors than of sound errors. In addition, they provide some evidence that word-sound-ambiguous errors like "Tar--- Car Talk", in which the error forms an existing word of English that differs from the target word by only one segment, should be regarded as sound-level errors despite their surface wordness. This observation raises an interesting question about the processing locus at which the correction behavior is determined. It might have been a reasonable hypothesis that the decision to mark the correction prosodically is made on the basis of the monitored output, i.e. if the error output produces a real word, which might mislead the listener, then it is more likely to received a marked correction to bring it to the listener's attention and fend off communication troubles. On this view, errors which result in non-word outputs are not as likely to disrupt the communication process, and so their corrections are less likely to be marked.

However, the results reported here suggest that some wrong words produced in error are likely to receive marked corrections (i.e. those in which target and error words are very different in their phonological shape), while others are not (i.e. those that are quite similar in their phonological shape.) How could this difference arise? One possibility is that the correction behavior is different because the locus of occurrence of the error is different. On this view, word-selection or -ordering processes might trigger a different monitoring and correcting mechanism from sound-selection and –ordering processes. A second possibility is that the correction behavior is different because the

speaker recognizes that the listener will need more help recovering from a word error when the error word contains no phonological segments which might provide a clue to the target word. On this view, when the error word differs from the target by only one sound, the speaker knows that the listener has some evidence for the nature of the target word even before the correction is uttered, and so is less likely to mark the correction prosodically.

The present results do not distinguish between these two possible accounts,but one way of doing so would be to determine whether the most common types of errors (e.g. interactions between word-onset consonants) are less likely to receive marked corrections than rarer types of error (e.g. whole-syllable exchanges.) Such a finding would support a model in which the correction behavior arises from the speaker's analysis of the surface output, and estimate of the listener's need for extra information. However, the observation that corrections of word-sound-ambiguous errors are similar to corrections of sound errors suggests that, despite their surface word shape, these errors arise as sound errors, and favors the alternative view that the correction difference is somehow related to the error mechanism rather than the wordness of the error outcome.

This view is also supported by another line of argument. Every natural human language has a vocabulary running into the tens of thousands, but no language has a phonemic inventory running into even the low hundreds. The largest phonemic inventory size listed by Maddieson [20] is 141, and the mean and the median in that database both lie around 30. English, with an inventory in the forties, is thus in the top quartile of the languages surveyed by Maddieson. Nonetheless, the size of the phonemic repertoire is obviously trivial in comparison to the size of the vocabulary. Not only are there few phonemes; strict constraints rule the order in which they may occur to constitute a word. Thus the string "string" contains five phonemes, but of the 120 orderings that are conceivably possible for a string of five elements, only one or possibly two ("string"itself, plus "ringst", rhyming with "jinxed") are allowed by the phonotactic constraints of English (a couple more are pronounceable, but are ruled out by voicing assimilation constraints). Inevitably, in such a situation, the words of a language resemble one another. And for similar resons, sound errors will strongly resemble real words. It has long been known that errors tend in general to conform to the phonotactic constraints of the language [1]; no speaker is likely to produce "rngtsi"for "string". Thus, errors only have freedom to select among the allowable strings of the language, and since the small number of phonemes combined with the large number of words means that many allowable strings are already in use, it is likewise inevitable that substitution or movement of a phonetic segment will run a very good chance of fortuitously producing an existing real word.

Thus we would claim that the considerable debate over ambiguity of errors like "Tar Talk" for "Car Talk" may be a red herring. Prosodic analysis suggests that these errors are sound-level errors which simply happen to have resulted in a string which is already in use in the vocabulary.

ACKNOWLEDGMENTS

We gratefully acknowledge the financial support of the NIH, (RO1-DC02125, DC02978) and the MIT Undergraduate Research Opportunities Program, as well as the generosity of Elizabeth Shriberg at SRI International in harvesting the errors from the ATIS corpus.

REFERENCES

[1] Fromkin, V.A. 1971. The non-anomalous nature of anomalous utterances. Language, 47, 27-52.

[2] Garrett, M.F. 1975. The analysis of sentence production. In G.H. Bower (ed.) The Psychology of Learning and Motivation. Vol. 9. New York: Academic Press.

[3] Garrett, M.F. 1988. Processes in language production. In F.J. Newmeyer (ed.) The Cambridge Linguistic Survey. Vol. III: Psychological and Biological Aspects of Language. Cambridge: Cambridge University Press; 69-96.

[3] Nooteboom, S.G. 1980. Speaking and unspeaking: Detection and correction of phonological errors in spontaneous speech. In V.A. Fromkin (ed.) Errors in Linguistic Performance: Slips of the Tongue, Ear, Pen and Hand. N.Y.: Academic Press. 87-95

[4] Dell, G.S. and Reich, P.A. 1981. Stages in sentence production: an analysis of speech error data. Journal of Verbal Learning and Verbal Behavior, 20, 611-629.

[5] Stemberger, J.P. 1985. An interactive activation model of language production. In A.W. Ellis (ed.) Progress in the Psychology of Language. Vol. I. London: Erlbaum.

[6] Shattuck-Hufnagel, S. 1992. The role of word structure in segmental serial ordering. Cognition, 42, 213-259

[7] Shattuck-Hufnagel, S. 1983. Sublexical units and suprasegmental structure in speech production planning. In P.F. MacNeilage (ed.) The Production of Speech. New York: Springer; 109-136.

[8] Levelt, W.J.M. 1989. Speaking. Cambridge, MA: MIT Press.

[9] Levelt, W.J.M., Roelofs, A. and Meyer, A.S. 1999. A theory of lexical access in speech production. Behavioral and Brain Sciences,

[10] Shattuck-Hufnagel, S. and Klatt, D.H. 1979. The limited use of distinctive features and markedness inspeech production:. JVLVB 18, 41-55

[11] Mackay. D. 1987. The organization of perception and action: A theory for language and other cognitive skills. New York: Springer

[12] Cutler, A. 1983. Speakers' conceptions of the functions of prosody. In A. Cutler and D.R. Ladd (eds.) Prosody: Models and Measurements. Heidelberg: Springer; 79-91.

[13] Deese, J. 1984. Thought into Speech: Psychology of a Language. Englewood Cliffs: Pretice Hall.

[14] Goffman, E. 1981. Radio talk. In E. Goffman (ed.) Forms of Talk. Oxford: Blackwell; 197-327.

[15] Levelt, W.J.M. and Cutler, A. 1983. Prosodic marking in speech repair. Journal of Semantics, 2, 205-217.

[16] Pierrehumbert, J.B. 1980. The phonology and phonetics of English intonation. MIT PhD thesis, distributed by the Indiana University Linguistics Club.

[17] Beckman, M. and Pierrehumbert, J.B. 1986. Intonational structure in Japanese and English. Phonology Yearbook, 3, 255-309

[18] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J. 1992. ToBI: A standard for labeling English prosody. In Proc. of the International Conference on Spoken Language Processing (ICSLP), Banff, v. II, 867-70

[19] Pitrelli, J., Beckman, M., and Hirschberg, J. 1994. Evaluation of prosodic transcription labeling reliability in the ToBI framework. In Proceedings of the International Conference on Spoken Language Processing (ICSLP), Yokohama, v. I, 123-126

[20] Maddieson, I. 1984. Patterns of Sounds. Cambridge: Cambridge Press.