Using pronunciation data as a starting point in modeling word recognition

Mark A. Pitt and Keith Johnson
Ohio State University, Columbus, OH
E-mail: pitt.2@osu.edu, kjohnson@ling.ohio-state.edu

ABSTRACT

Many of the mysteries of spoken word recognition have evolved out of the observation that pronunciation is highly variable yet perception is amazingly stable (i.e., listeners perceive the intended word). Proposed solutions to this perceptual constancy problem tend to be process oriented, in which mental processes restore or recover the intended word form en route to lexical memory, or representation -oriented, in which the variation itself forms part of the word's lexical entry. The viability of both approaches was examined by studying the acoustic and phonological variability found in the ViC corpus of conversational speech. The results highlight obstacles that they must overcome.

INTRODUCTION

Language databases, from online dictionaries to fully annotated corpora, are a mainstay of scientists interested in language processing, whether by man or machine. They provide a window into how language is used, both written and spoken, serving as a rich source of information with which to test hypotheses and develop insights into linguistic communication. The present paper follows in this vein, using a corpus of conversational speech to examine proposals on how spoken words are recognized.

Phonological variation, particularly regressive assimilation, has begun to attract the interest of psycholinguists. Like the extensive body of work on acoustic-phonetic variability in speech perception, the task is to understand how perceptual stability is achieved amidst production variation, only in this case the variation is extreme enough to yield different phonetic percepts (e.g., *green ball* → *greem ball*).

Two broad classes of models have been proposed to explain how the listener perceives the assimilated variant as the intended word (e.g.,*greem* as *green*). One idea is that a phonological inference process recovers the underlying form of the segment when an assimilated word is encountered [1-3]. Processing-based proposals like this are attractive and intuitive because a regularity cross production environments is captured in a simple rule that has the potential to improve processing accuracy and efficiency at minimal cost.

Representation-based models offer an alternative approach to recognizing assimilated variants. The task of variant recognition is shifted to the lexicon. In her FUL (Featurally Underspecified Lexicon) model, Lahiri proposed that lexical representations are underspecified for features that cause nuisance variation such as assimilation [4]. Words are represented as sets of phonological features except that the place-of-articulation feature [CORONAL] is not specified (i.e., unmarked), which functionally makes the model insensitive to variation in this feature. Recognition of an assimilated variant is not a problem because the assimilation is not detected.

Lahiri's abstractionist approach stands in sharp contrast to instance (exemplar) models, in which the nuances of production variation are considered useful information to aid recognition [5,6]. Production variation such as assimilation is encoded in the lexical representation of the word, thereby providing a history of a word's realization that can be referenced during lexical matching process.

Our aim in this project was to evaluate the preceding accounts of recognizing assimilated variants by studying assimilation itself. Little is known about assimilation in informal speech, so knowledge of its realization should assist in evaluating the proposed solutions. A clear understanding of the challenge assimilation posses to modeling its recognition requires studying production variation in the environments in which assimilation could occur and should not occur. To this end, we carried out phonological and acoustic analyses of assimilation in a corpus of conversational speech. The data define the phenomenon to be modeled and are used to evaluate the models.

The Buckeye corpus of conversational American English served as the source of speech for this investigation [7]. It was collected using a modified interview format and comprises 300,000 words from 40 talkers born and raised in central Ohio. To date, one third of the corpus (14 talkers) has been phonetically transcribed and was used in the following analyses.

PHONOLOGICAL ANALYSES

Phonological analyses were restricted to three word-final coronals, [n], [d],and [t], followed by a word-initial labial or dorsal that contrasted in place of articulation ([m], [b], [p], [g], [k]). An environment that permits assimilation (e.g., *green ball*) occurred 21% of time in the corpus, which is surprisingly high and suggests that assimilation could be quite a frequent form of variation. Despite this potential, assimilation occurred only 10% of time in this environment. To understand why it did not occur more often, as well as what other forms of variation occurred, we tallied the frequency of all forms of variation found with the three coronals (Figure 1).

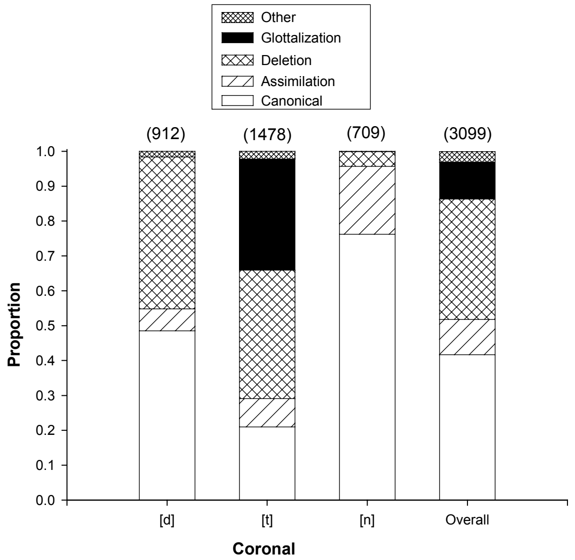

Figure 1. Frequency of various realizations of [ndt]. Numbers in parentheses specify the total number of word tokens available for analysis.

Overall, deletion of the coronal was more than three times as likely to occur than its assimilation, but there was also variability across segments. [d] and [t] underwent deletion and assimilation to similar degrees, but [d] was never glottalized and [t] was (0.30), which may account for why canonical realizations of [d] were more than twice as frequent as [t]. [n], on the other hand, never deleted, which may explain why it assimilated more than twice as often as the two stops (0.20 vs. 0.07).

These data indicate that assimilation is a minor form of variation because the canonical form of the phone and its deletion dominate in production. The much higher frequency of [d] and [t] deletion also suggests that any account of how assimilated variants are recognized should be extendable to include these more common realizations, which occur not just in environments where assimilation is possible, but in many others as well [8].

Even if the issue of generalizability to other forms of variation is ignored, corpus analyses also revealed that significant issues remain even when the problem is restricted to assimilation. For example, any proposal must be able to discriminate environments in which assimilation occurred (e.g., *greem ball*) from those in which it did not (e.g., *team ball*). This is a formidable problem because both occur about equally often (10% vs. 9%, respectively). A simple decision rule is bound to make errors (e.g., false alarms or misses). Analyses also showed that assimilation is not limited to following dorsal and labial contexts. A small percentage of time (0.5%) assimilation was found before glides and liquids (e.g., *saib well*; *coulb really*). Most surprisingly of all, 3% of the time assimilation occurred in the reverse direction, from labial to coronal (e.g., *I'n doing*). In another 3% of instances, the nasal in a word-final cluster (e.g.,[nt]) assimilated to the following place of articulation when [t] deleted (e.g., *innocem people*). Although each of these instances is somewhat infrequent, when combined they are too numerous to be ignored.

The results of the preceding analyses identify some of the complexities of assimilation. In particular, assimilation is not as context-specific as one might suppose and identification of a true instance of assimilation is likely to be nontrivial. Although the present data can assist in modeling the phenomenon, given the rarity of assimilation and the frequency of other forms of variation (e.g., deletion), we believe that it is most productive to consider solutions that are sufficiently general to apply to all forms of production variation. A purely processing-based model may be ill-suited to the task. In the case of assimilation, any inference mechanism would need to be quite sophisticated to perform accurately across all of the contexts discussed above. In the case of deletion, it is unclear how a deleted segment (feature or phoneme) would be identified and what event in the speech stream would trigger restoration of the missing segment, let alone the proper segment.

A representation-based model, of course, faces the same challenges, but a lexically-based solution seems better-suited to the general problem of phonological variation, all forms of which are treated similarly: They are a problem of matching the speech input onto lexical memory. Assimilations do not have to be undone. Rather, such variation merely affects the quality of the match, although the exact details differ between models (see above). Deleted segments do not have to be restored. As with assimilation, they would do little more than affect the match. The brains of this type of model lie in the decision mechanism that governs lexical selection of the intended word from amongst competitors. Its accuracy will be a major factor in determining the success of the approach.

ACOUSTIC ANALYSES

We examined the acoustics of assimilated stops (not nasals) to determine whether residual information about the underlying segment remains in the signal even in tokens that are transcribed as assimilated. Several researchers have reported that phonological processes that eliminate lexical contrast or obscure lexical identity when examined in phonetic transcription exhibit fine-grained acoustic information that might be useful to listeners [9-14]. Our analyses are an extension of this work to the case of assimilation in conversational speech.

Measurements of assimilated oral and nasal stops were compared with those of unassimilated dorsal, coronal, and labial stops. To insure that the unassimilated stops were truly unassimilated, we examined only consonants that were followed by a consonant having the same place of articulation, such as job before [labial], went down [coronal], and lack good [dorsal]. The assimilated were underlyingly [coronal] but had been transcribed as either [labial] (or [dorsal]), and preceded a conditioning [labial] (or [dorsal])..

The trajectory of the second formant was examined in the vowel preceding each stop consonant. Vowel formant frequencies were measured in two locations - at the vowel midpoint and just before the consonant closure. The formants were measured automatically using an implementation of Robust LPC [15], and were hand corrected afterwards. The window size of the analysis was kept short (10 ms) so that the rapidly changing part of the vowel just before the consonant closure would be measured.

Three vowels had enough tokens in each category to permit comparison of vowel formants for each of the [labial], [coronal], [dorsal], [labial]<-[coronal], and [dorsal]<-[coronal] conditions. These were /I/, /ɛ/, and /a/. Only the results for /I/ are shown (Figure 2), as those for the other two vowels are quite similar. The F2 trajectory is represented in these figures as the difference in hertz between the F2 at the center of the vowel and the F2 at the onset of the consonant so that when F2 falls from the vowel to the consonant (as is usually the case for [labial] consonants) the F2 difference is a large positive value, and when the F2 rises from the vowel to the consonant (as is usually the case with [dorsal]) the F2 difference is a large negative number. The assimilated consonants have a formant trajectory distribution that falls in between the formant trajectories of the underlying and surface forms. There are some tokens of [b]←/d/ that have a typical F2 trajectory and some that fall squarely in the range for /d/. The median formant trajectories though for assimilated consonants do not match the median value for either the underlying source or the supposed target of the assimilation. Many tokens fall in intermediate regions on the F2 difference scale.

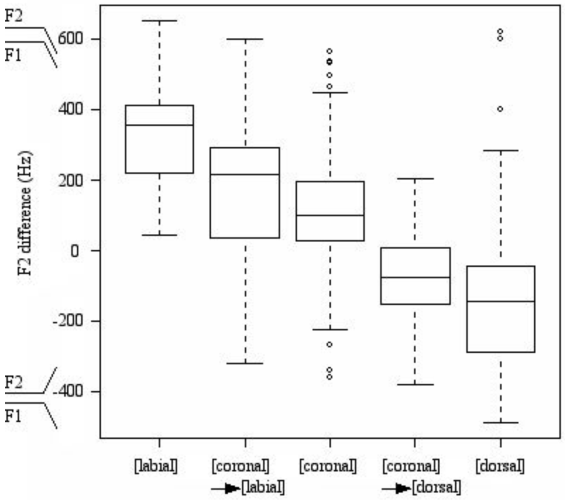

Figure 2. A boxplot showing the distributions of F2 difference (F2 vowel minus F2 consonant) measurements for assimilated and unassimilated stops and nasals (pooled) following /I/. The column labeled [labial] includes items that ended in /b/,/p/ or /m/ and were followed by a [labial]. The columns labeled [coronal] and [dorsal] are analogous. The column labeled [coronal]→[labial] contains words that ended in /d/ ,/t/ or /n/ underlyingly but were realized as [b],[p], or [m], respectively, in the context of a following [labial]. Similarly, the column labeled [coronal]→[dorsal] includes items that were underlying [coronal] but realized as [dorsal] with a following [dorsal] consonant. Stylized F1,F2 trajectories on the vertical axis show trajectory shapes that are associated with the F2 difference scale.

Two examples of F2 variation in these environments are shown in Figure 3. The frequency of F2 is plotted over time for matched triples of word pairs in which the assimilated and nonassimilated tokens were produced by the same talker and were preceded by the same vowel. The top graph shows a case of labial assimilation. The /t/ of got is pronounced with lip closure before the following [m] - *gop my*. The trajectory of F2 in cop pulls shows the realization of an underlying [labial] in this vowel environment, and the trajectory of F2 in got pulled shows an unassimilated /t/. The trajectory of assimilated gop is in between the trajectories of the two unassimilated consonants. The bottom panel shows a similar set of tokens for the dorsal assimilation closet gay in which the final /t/ was realized as [k]. Again, the assimilated token shows and F2 trajectory that is intermediate between unassimilated /k/ and /t/.

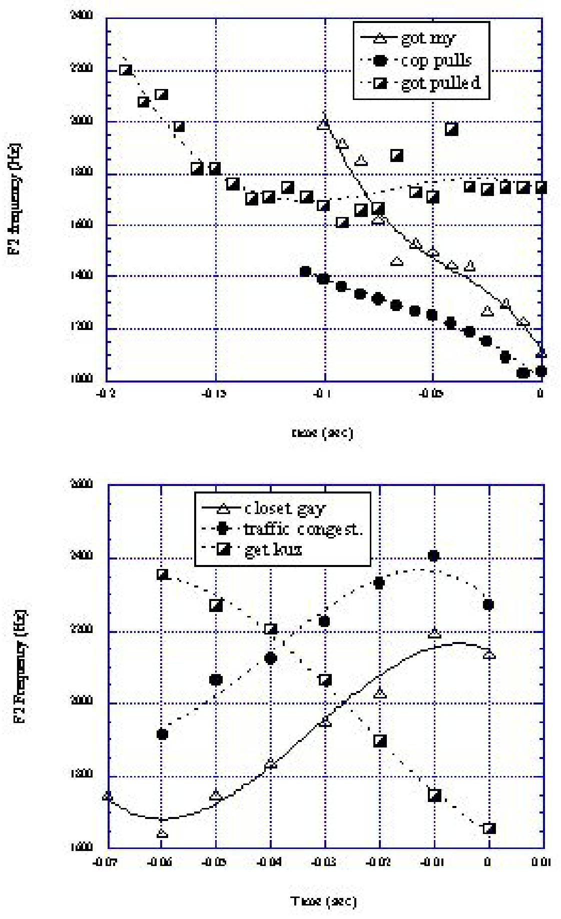

Figure 3. F2 trajectories of examples of assimilated and nonassimilated tokens in which the talker and vowel identity are matched. The assimilated tokens *gop(t) my* and *closek(t) gay* are plotted with open triangles.

These data indicate that place assimilation does not obliterate place of articulation information in the acoustic signal. On the contrary, even in conversational speech residual information about the assimilated consonant is present much of the time.

Processing-based models could use this information to improve the accuracy of the inference about the underlying segment. An instance model could do the equivalent by encoding production variation in the lexical entry of the intended word. FUL, in contrast, was explicitly developed to be insensitive to such variation so recognition would not be misled by it. Although this design strategy has obvious advantages, it is unclear how the recognition system would selectively learn to ignore the acoustic information that specifies unmarked segments, such as place of articulation of coronals. The F2 distributions of the assimilated stops are just another example of the phonetic ambiguity that is ubiquitous in speech. If the acoustic variation of marked and unmarked segments is not qualitatively different, then it may be very difficult indeed to build a model that uses only the former and ignores the latter.

CONCLUSION

The data that we present here on place assimilation in conversational speech provide new details about the phenomenon, which in turn provide new challenges for models of recognition. The results of the phonological analyses suggest that a model must be capable of distinguishing true assimilations from nonassimilation (e.g., *team ball*). Just as importantly, a model must be able to recognize the (more frequent) deleted forms of the words. The acoustic analyses probed assimilation further by focusing on the cues that specify stop identity. Acoustic information often rendered assimilated segments as acoustically distinct from nonassimilated segments that had the same phonetic transcription.

Corpus analyses provide a detailed and comprehensive description of the speech input to the recognition system. To be taken seriously, a model must be able to simulate human recognition given this input, either computationally or through a detailed explanation. Those that fail to pass this sufficiency test must be modified or eliminated. Ultimately, experimental data that reveal the in workings of the recognition system will be needed to choose between models, but corpus data can help in narrowing the field by establishing minimal performance criteria that any model must meet to be considered a plausible account of the phenomenon.

The volume of data that corpus analyses can provide serves as a useful starting point for modeling recognition because it forces modelers to be quite explicit about how recognition is accomplished given an accurate description of the input to the recognition system. Precise predictions can then be derived from the model and tested in laboratory experiments. Used in this way, the two approaches are sure to get us closer to the goal of understanding how spoken words are recognized.

REFERENCES

[1] Deelman, T., & Connine, C. M. (2001). Missing information in spoken word recognition: Nonreleased stop consonants. *Journal of Experimental Psychology: Human Perception & Performance, 2*7, 656-663.

[2] Gaskell. G., & Marslen-Wilson. W.D. (1996). Phonological variation and inference in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 144-158.

[3] Gaskell. G., & Marslen-Wilson. W.D. (1998). Mechanisms of phonological inference in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 380-396.

[4] Lahiri, A. (1999). Speech recognition with phonological features. *The XIV International Congress of Phonetic Sciences,* San Francisco.

[5] Johnson, K. (1997a) Speech perception without speaker normalization. In K. Johnson and J.W. Mullennix (Eds) *Talker Variability in Speech Processing* (pp. 145-166) NY: Academic Press.

[6] Johnson, K. (1997b) The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Linguistics* 50, 101-113.

[7] Pitt, Mark, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. (2003). *The ViC corpus of conversational speech.*

[8] Raymond, William D., Robin Dautricourt, and Elizabeth Hume. (2003). *Medial /t,d/ deletion in spontaneous speech.*

[9] Barry, M.C. (1985) A palatographic study of connected speech processes. *Cambridge papers in phonetics and experimental linguistics* (vol. 4, pp. 1-16). Cambridge, England: Department of Linguistics, University of Cambridge.

[10] Charles-Luce, J. (1997) Cognitive factors involved in preserving a phonemic contrast, *Language and Speech, 40*, 229-248.

[11] Ellis, L. & Hardcastle, W.J. (2002). Categorical and gradient properties of assimilation in alveolar to velar sequences: Evidence from EPG and EMA data. Journal of *Phonetics, 30*, 373-396.

[12] Kerswill, P.E. (1985) A sociophonetic study of connected speech processes in Cambridge English: An outline and some results. *Cambridge papers in phonetics and experimental linguistics* (vol. 4, pp. 17-39). Cambridge, England: Department of Linguistics, University of Cambridge.

[13] Manuel, S.Y. (1991). Recovery of "Deleted" Schwa. In O.Engstrand & C. Kylander (Eds.). *PERILUS XIV*, 115-118.

[14] Manuel, S.Y., Shattuck-Hufnagel, S., Huffman, M., Stevens, K.N., Carlson, R., & Hunnicut, S. (1992). Studies of vowel and consonant reduction. In J.J. Ohala, T.M. Nearey, B.L., Derwing, M.M. Hodge & G.E. Wiebe (Eds.). Proceedings of the 1992 *International Conference on Spoken Language Processing*, 943-946.

[15] Lee, C-H (1988) On robust linear prediction of speech. *IEEE Transactions on Acoustics, Speech, and Signal Processing 36,* 642-650.