PATTERN RECOGNITION IN SPEECH PERCEPTION RESEARCH

Terrance M. Nearey
Dept. of Linguistics, University of Alberta, Edmonton AB, Canada T6G 2E7
email t.nearey@ualberta.ca

1. INTRODUCTION

We have successfully applied simple pattern recognition techniques to several problems of speech perception. Nearey and Hogan (1986) describe the NAPP (normal *a posteriori* probability) method, based on linear discriminant analysis. Nearey and Assmann (1986) apply NAPP to accurately predict listeners' behavior in the perception of modified natural English vowels. Recently, variations of NAPP have also been applied successfully to cross-linguistic and L2 vowel perception (Morrison 2006, Thompson 2007). Direct application of NAPP involves training a pattern recognizer on natural production measurements and using the 'frozen' model to predict listeners' behavior on new stimuli, without any further tuning. This paper sketches the use of more flexible pattern recognition methods in speech perception research. These include logistic regression and methods imported from automatic speech recognition (ASR) technology.

2. NAPP AND MNLR

The APP (*a posteriori* probability) scores generated by a NAPP model can be expressed in the form of a multinomial logistic regression (MNLR). MLNR has a long history in econometrics (Train 2003) to model discrete choice (by, e.g., consumers). MNLR can be tuned to approximate listeners' response data directly. The question of phonetic compositionality of perceptual choices among (e.g.) CV or VC syllables has been investigated extensively by Nearey (e.g., 1997) using MNLR techniques. For the cases studied, listeners' sensitivity to stimulus properties seems to be linked phoneme- or subphoneme-sized units. Larger units such as syllables, do not appear to associate with specific stimulus patterns in the ways that lower level units do. A related application is discussed below.

3. VC(C)V SYLLABLES

Nearey and Smits (2002) describe a variation of an experiment by Repp (1983) which involves variable (phoneme) length utterances of the form of VC(C)V. We were not at all clear that MNLR models would reveal the kind of compositionality we found with simpler response sets. Our experiment spanned the following responses {aba, ada, ab#ba, ad#da, ab.da and ab.da}, where '#' indicates a phonotactically necessary (in English) word boundary and '.' Indicates a syllable (and possibly word) boundary. The vowel denoted 'a' is low back and slightly rounded in the dialect under study.

3.1 Method

A total of 144 (6 x 6 x 4) stimuli were created by a standard Klatt80 synthesizer. The stimuli were arrayed in *fully crossed* 3-factor design with the following values: 1) *Closing F2* (and correlated F3) associated with VC (ab- and ad-) offset [1060 (2180) to 1450 (2539) Hz in 6 steps]. 2) *Opening F2* (and correlated F3) with CV (-ba and –da) onset [1099 (2262) to 1635 (2500) Hz in 6 steps]. 3) *Gap Duration* at 4 levels 80, 120, 190 and 300 ms. The initial vocoid, V1 had [F1 F2 F3] targets of [777 1147 2466] Hz and a fixed duration of 190 ms including 50 ms V1C transitions; The final vocoid, V2, had the same target frequencies and a duration 300 ms including the CV2 transitions, F0 was fixed at 125 Hz for V1. And a linear downward trend from 125 to 100 Hz for V2. The amplitude of voicing (av) was set to 60 dB at the beginning of V1, it fell abruptly to 0 dB at V1 offset and rose abruptly to 60 dB at V2 onset. Participants were 13 native speakers of Canadian English. Each responded to 10 repetitions of each of the 144 stimuli. Response button layout on a PC screen was as follows:

[b] [bb] [bd]
[d] [dd] [db]

3.2 Results

An initial MNLR analysis was conducted (Nearey and Smits 2002). The response factor comprised the 6 response categories above. The independent variables were the *Closing F2*, *Opening F2* and *Gap Duration* (expressed as square root of ms). This model provided a relatively good fit, with a residual RMS error of about 6 percentage points. The model predicted the modal (most popular) response of listeners for almost 94% (135/144) of the stimuli.

The clustering patterns of consonant responses in Figure 1 (and associated t-tests) suggested a factored (compositional) solution, whereby judgment log-odds were tuned continuously by only the following three factors: (1) Closing place (ab-/ad-) was tuned only by *Closing F2;* (2) opening place (-ba-/da-) only by *Opening F2.* Finally, define a third phonetic factor, cluster type, comprising singletons (-b-, -d-), geminates (-b#b-, -d#d-) and true clusters (-b.d-, -d.b-). Then cluster type is tuned by (3) *Gap Duration* only. A reduced logistic model enforcing the decomposition above shows RMS error and modal agreement that are nearly indistinguishable from the full CC model. Note that this analysis involves splitting even singleton stops (e.g., -b-) into *closing*, *gap*, and *opening* subparts that can be shared with other C(C) patterns.

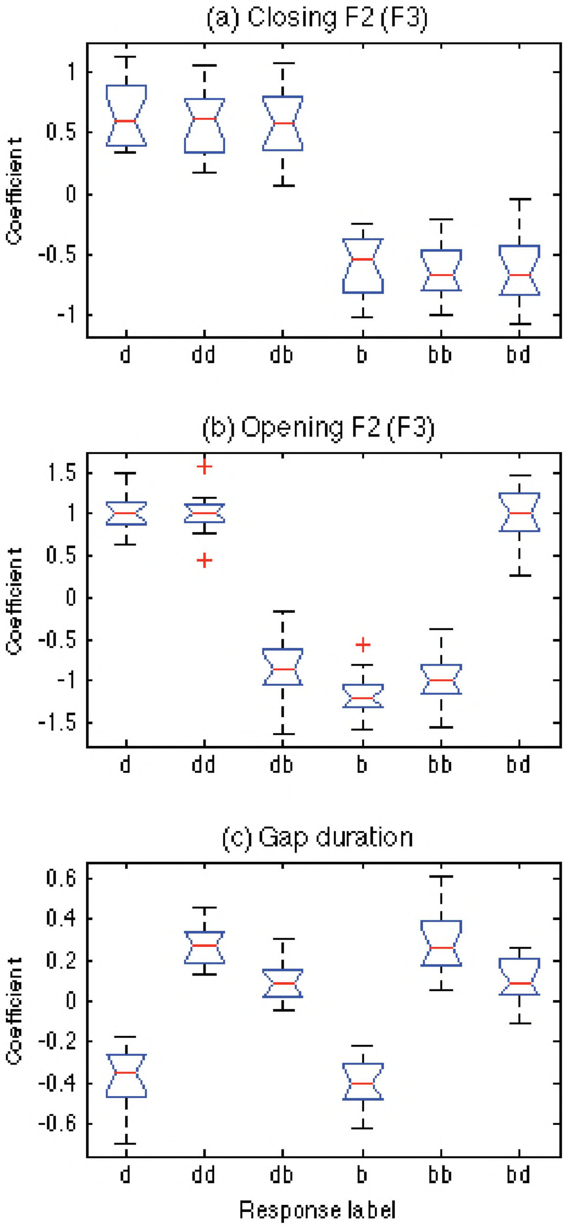

Fig. 1. Coefficients of logistic regression for stimulus properties in experiment 1.

4. BRIDGING TO ASR METHODS

The methods describes so far involve *static* pattern recognition, requiring exactly the same number of signal properties for each stimulus. Outside the laboratory, speech chunks come in many sizes. Longer ones have more properties than shorter ones (e.g., *Albuquerque* vs. *Al*). ASR technology has developed several *dynamic* pattern recognition methods that handle such variable inputs. Preliminary experiments (Nearey 2004) match the results of section 3 using a (dynamic) hidden semi-Markov model (HSMM, Guédon, 1992). The model uses sub-phone states (e.g, *b-closing, cluster-gap, d-opening*) directly related to the subphone elements of section 3. Using a maximum mutual information criterion, the HSMM can be tuned to listeners' responses to performs as well as the MNLR models above. The HSMM uses a conventional frame-based, mel frequency cepstrum representation The resulting system provides a complete framework for modeling speech perception that starts with raw waveforms and culminates in accurate predictions of listeners' responses. This first step bodes well for the future of incorporation of modeling technologies from ASR directly into speech perception research. With time, it may facilitate feedback in the other direction.

REFERENCES

Guédon, Y. (1992). Review of several stochastic speech unit models. Comput. Speech and Lang., 6(4), 377-402.

Nearey, T. M. (1997). Speech perception as pattern recognition. J. Acoust. Soc. Amer., 101(6), 3241--3256.

Nearey, T. M., & Hogan, J. (1986). Phonological contrast in experimental phonetics In J. Ohala, & J. Jaeger (Eds.), Experimental Phonology (pp. 141-161.). New York: Academic Press.

Nearey, T. M., & Smits, R. (2002). Patterns in the Perception of VC(C)V strings. J. Acoust. Soc. Amer.,, 112, 2323 (abstract).

Nearey, T. M. (2004). Adapting automatic speech recognition methods to speech perception: A hidden semi-Markov model of listener's categorization of a VC (C) V continuum. J. Acoust. Soc. Amer., 116, 2570 (abstract).

Repp, B. H. (1983). Bidirectional contrast effects in the perception of VC-CV sequences. Perception and Psychophysics, 33(2), 147-155.

Train, K. (2003). Discrete Choice Methods with Simulation. Cambridge: Cambridge University Press.

ACKNOWLEDGMENTS

This research was funded by SSHRC.