IMPLICIT RATE AND SPEAKER NORMALIZATION IN A CONTEXT-RICH PHONETIC EXEMPLAR MODEL

Travis Wade

Institute of Natural Language Processing, University of Stuttgart, Germany
travis.wade@ims.uni-stuttgart.de

ABSTRACT

In this study we present a model of speech perception in which (1) memory includes a single, ordered collection of acoustic cues extracted at landmark locations from previously heard signals and encoded to preserve temporal patterns, and (2) identification of newly encountered sounds involves comparing the sounds—and their surrounding contexts—with similar sequences occurring in memory. Under these assumptions, perceptual speaker and rate normalization and context dependence in general follow implicitly from the statistics of the language environment and do not require traditionally assumed processes or levels of representation. We verify this by means of a simulation in which the model simultaneously acquires VOT and F1 cues to consonant voicing and vowel height, and their dependence on speaking rate and speaker gender, based on exposure to productions from the TIMIT database.

1. INTRODUCTION

Exemplar approaches to phonetic perception, acquisition, memory and evolution emphasize the role of high-order statistical regularities that occur over large numbers of perceived (or produced) utterances, and question the need for abstract rules, processes, and structures that have traditionally been used to explain these regularities [e.g. 1, 2, 3, 4]. One key prediction of such an approach is that a memory of richly specified phonetic representations can result in automatic, implicit perceptual compensation for the effects of context (segmental, speaker, rate, etc.) on the acoustic realizations of speech sounds. If perception involves comparing a newly encountered sound with actual previously identified sounds, and if some of the dimensions along which this comparison is made include information about the context in which the sound occurred, then patterns of covariance between contextual and more "primary" cues are preserved in the identification process.

Of course, this preservation depends on informative contextual cues being (1) sufficient and salient in the speech signal, (2) appropriately represented in memory, and (3) allowed to contribute to relevant comparisons. Although several quantitative phonetic exemplar models have been proposed [1, 2, 5], incorporation of context information so far is rarer and more speculative, especially where the context is distributed over time. In this paper, we introduce a unified model of context specification and use it to test whether acoustic cue distributions observed in speech production are sufficient to account for known perceptual compensation patterns. Specifically, we show that the temporal co-occurrence of different cues to consonant voicing and vowel height in the TIMIT database trivially predicts "normalization" for gender and speaking rate in a model that references neither of these variables directly.

In our model, there is no explicit structural analysis or segmentation of incoming speech. Memory consists simply of an ordered collection of richly specified, real-time descriptions of perceived sounds, not unlike a continuous recording of years of auditory input. This sequence is sparsely annotated with a record of other events that co-occurred with the speech, including analyses or categorizations. Perception, then, involves identification of annotations occurring near regions of memory that are similar to a (similarly specified) input pattern. In the present study, we assume that the acoustic descriptions consist of potentially informative parameter values extracted at salient landmarks in the speech signal [6]. For simplicity, we further assume (see discussion) that the labels correspond roughly to traditionally described distinctive features. Whenever landmarks are detected, the relevant cues are marked at corresponding locations in a continuous memory "signal". During perception, a newly encountered sequence of landmarks "resonates" with similar regions of the memory, causing nearby labels to be activated. In the next

sections, we discuss context effects in a set of production data, and describe in detail how the data are stored and compared in the model.

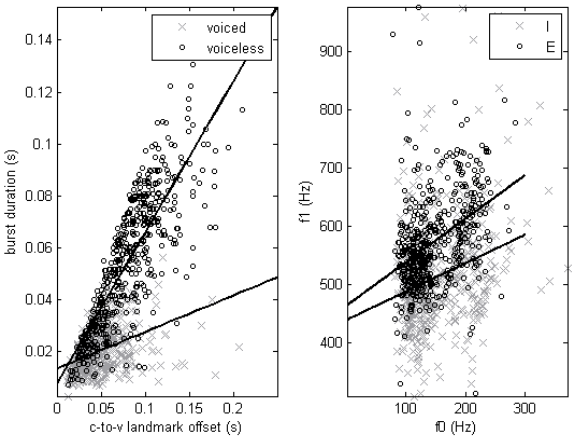
2. SPEECH DATA

It is well known that first formant frequency is a primary indicator of vowel height, and that voice onset time is related to consonant voicing. However, neither cue (or any other) is sufficient to classify naturally occurring sounds, since F1 and VOT are also affected by (e.g.) speaker gender and speaking rate, respectively. Fortunately, humans have little difficulty compensating for these influences and make F1-based [height] and VOT-based [voiced] distinctions appropriate for a particular context [7, 8].

The acoustic data we considered were VOT, f0 and F1 values measured at consonant and vowel landmarks in the context of different rates and speakers. Specifically, we analyzed all of the sequences of stop consonants followed by the vowels [I] and [ɛ] in the training portion of the TIMIT database [9], across syllabic/prosodic status and speaker gender and dialect, a total of 1824 tokens. Consonant landmarks were taken to be the first sample of a [b], [d], [g], [p], [t], or [k] burst as labeled in the database. Vowel landmarks were taken to be maxima in the envelope of the sound below 500 Hz during the regions labeled as consonant and vowel segments (the CV plus 50ms linear-ramped precursor and following context was low-pass filtered at 500 Hz, full-wave rectified, and then low-passed at 40 Hz). VOT was simply taken as the duration of the consonant release burst labeled in the corpus. F0 was derived using an autocorrelation-based algorithm, and F1 using the Burg algorithm (both as implemented in [11]) to find five formants below 5000 Hz (male speakers) or 5500 Hz (females). F0 and F1 values were averaged over a 40 ms window around the vowel landmark. Exemplars where pitch could not be estimated or with F1 estimates above 1000 Hz (<1%) were discarded. Finally, the dataset was trimmed such that the 4 CV sequence types ([+voiced], [+high]; [+voiced], [-high]; [-voiced], [+high]; [-voiced], [-high]) occurred in equal numbers and such that the average consonant-to-vowel landmark offset was as nearly equal as possible between voiced and voiceless stimuli. This was to minimize effects of coincidental a priori frequency of occurrence or average speaking rate on categorization. The result was a training set of 1000 exemplars, 250 for each [voiced]/[high] combination.

Figure 1 shows the resulting distributions. As might be expected, VOT was influenced by consonant voicing, but also correlated with speaking rate—as reflected (indirectly) in the time between measured consonant and vowel landmark locations—for voiceless consonants. Similarly, F1 was influenced by both vowel height and f0. Even in the 2-dimensional spaces there was a great deal of overlap for both distinctions, presumably related to prosodic, speaker, and dialect differences and other influences.

Figure 1: Distributions of training stimuli in rate/VOT and f0/F1 space, best-fit lines for each stimulus type.



2.1. Training stimuli

Input to the model was a 3-dimensional sequence containing the acoustic measurements and annotated with feature labels. First, measured VOT, f0, and F1 values were normalized to a mean of 0, standard deviation 1. For each CV sequence, a point process was generated for each dimension that consisted of zeros everywhere except at the relevant landmark location, where it took the normalized parameter value. These sequences were sampled at 200 Hz and padded with 130 ms of silence.

Random encoding error was then introduced by distributing acoustic cue information over time, smoothing each sequence with a Gaussian filter (s.d. 10 ms). Stimuli were finally normalized to a total power of 1.0 and concatenated to the end of the memory sequence. [high] and [voiced] values (1 or 0) were marked at the center of a CV sequence. Sample stimuli can be seen in the top panels of Figure 2. Although we will not make any claims about the biological

likelihood of such a representation, it is worth noting that related methods are often used in describing correlated neural activity, and that learning rules have been proposed for similar temporal patterns [e.g. 10].

The training set consisted of the 1000 exemplars shown in Figure 1, added to memory in a pseudo-random order. No assumptions regarding attention or cue-weighting were made in the model; vowel and consonant landmarks contributed equally to the representation.

2.2. Test stimuli

Test stimuli were newly generated sequences composed in the same manner, to represent VOT and F1 continua in different rate and speaker contexts, respectively.

For voicing continua, VOT varied from 0 ms to 80 ms in 11 steps, in the context of either a 50 ms (fast speech) or 120 ms (slow speech) consonant-to-vowel landmark offset. F1 and f0 values were taken at random from the training set in 30 separate continuum categorization runs, so that the results did not depend on a particular vowel configuration.

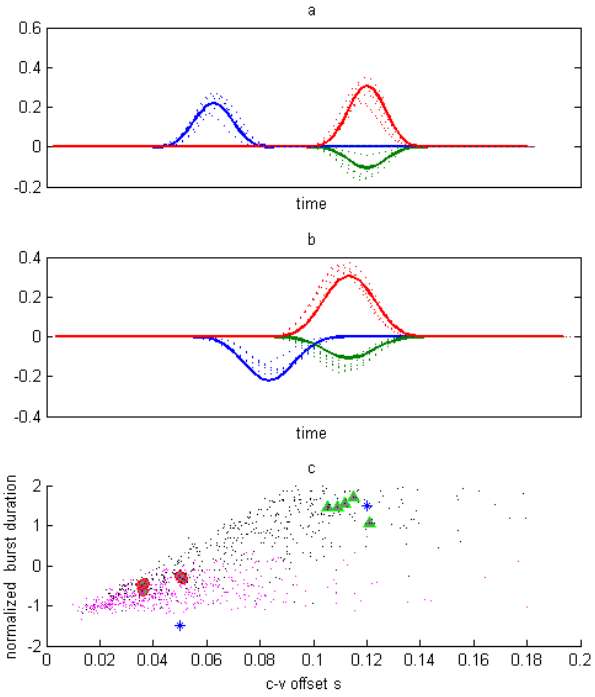
For vowel height continua, F1 varied from 340-750 Hz, in the context of either a typical male (-0.8 normed value) or a typical female (0.8 normed value) f0. VOT and landmark offset were taken at random from the training set in 30 separate runs, so that results did not depend on a particular timing configuration.

3. MODEL AND PROCEDURE

Categorization by the model consisted of identifying feature labels near peaks in the summed cross-correlation function of a test stimulus with the memory sequence. The five local peaks (i.e. $x_{i-1} < x_i > x_{i+1}$) with the highest absolute values were selected, and the average of the nearest [voiced] and [high] labels rounded (to 1 or 0) to achieve a classification in each dimension. An example of this process is shown in Figure 2. As demonstrated in the figure, sequences match better the more similar they are in both timing and (all of) the parameter values. Specifically, timing-related match is related to the asynchrony of two landmarks by a Gaussian function that is determined by the width of the filter described in section 2.1. Critically, however, no explicit rate or speaker information is involved in the representation or comparison.

Both VOT and F1 continuum stimuli were presented to the model in this way, and classification was averaged over the 30 runs for each condition described above.

Figure 2: Slow (a) and fast (b) stimulus sequences (solid lines), memory sequences corresponding to the top 5 matches (dotted lines); and (c) location of test (asterisks) and memory (circles and triangles) stimuli in vowel-landmark-offset by VOT space (as in Fig. 1).

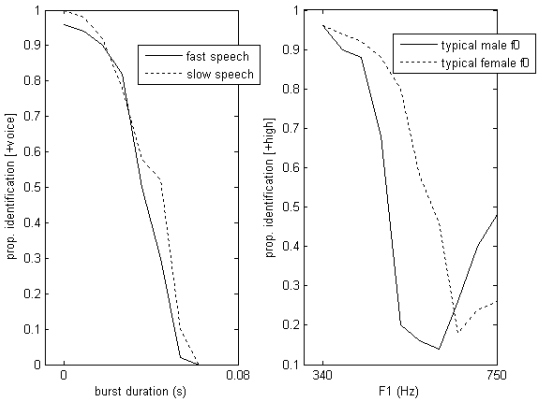


4. RESULTS

Average categorization of stimulus continua is shown in Figure 3. First, the monotonic decrease in [+voiced] and [+high] responses with increasing VOT and F1 indicates that, despite rate, gender, and dialect variability and without explicit segmentation the TIMIT data were sufficient to teach the model the importance of these cues (or, more precisely, that lower VOTs are likely to co-occur with voiced consonant labels, and lower F1s with high vowels.) A possible exception was vowels with very high F1 values produced by male speakers, where (see Figure 1) there was sparse, probably unreliable training data. In addition, comparison of overall mean categorization across conditions indicated that both rate and speaker normalization effects occurred. On average, the model accepted lower VOTs as signifying voiceless consonants at faster rates ($p=0.0015$), and higher F1s as representing high vowels in higher f0 contexts ($p=0.0017$). Importantly, these patterns

both mirror previously observed human context "normalization" effects [7, 8].

Figure 3: Classification of test stimuli across speaking rate (left) and speaker gender (right) contexts.



5. DISCUSSION

We have presented a quantitative model in which speech is remembered and perceived in context, without explicit normalization, segmentation, or unit categorization. It is often assumed that at least rate normalization is needed in order to avoid unmanageably large memory demand. We question this need, since exemplar models should predict rich memory specification in temporal as well as other dimensions, and in particular because we were able to model compensation for rate without normalization using a fairly small database of naturally occurring, highly variable productions.

The model described here is clearly incomplete, for several reasons. First, the variability shown in Figure 1 indicates that more than three dimensions will be needed to classify even the few sounds we considered across contexts, dialects, and speakers. Second, we made assumptions about representations in the model that need to be assessed empirically. In particular, we assumed that listener identifications and category markers in memory can be described in terms of traditionally assumed abstract, discrete features like [voiced] and [high]. We intended this representation as a shorthand for higher-order combinations of linguistic (perhaps word-, phrase-, or utterance level) and non-linguistic events or analyses, probably differing by listener and situation [e.g. 3, 12] that might have accompanied the sound sequence in memory. However, it may be that this simplification was not valid, and thus that we inappropriately imported assumptions about units from traditional theory. Similarly, the model assumes that speech signals are actually perceived and remembered as sequences of cues extracted at landmark locations. Human perception might not take the form of actively tracking, for example, F1 frequency. However, we assert that sets of cues like these are probably at least linearly related to what might be a more likely process, for example surprise detectors tuned to different frequency ranges, features, or parameters [2]. Finally, we did not include any reference to time decay of memory or attentional modulation in perception. We consider the present simulations to be conservative in not requiring these parameters, which could be straightforwardly included based on empirical data.

In sum, our results call into question some of the processes and structures that are needed to describe speech perception and memory. We suggest that exemplar models may go even further than is often assumed in accounting for phenomena traditionally ascribed to explicit normalization.

6. REFERENCES

[1] Pierrehumbert, J. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee and P. Hopper (eds.), *Frequency and the Emergence of Linguistic Structure*. Amsterdam: Benjamins, 137-157.

[2] Johnson, K. 1997. Speech perception without speaker normalization: An exemplar model. In K. Johnson and K. Mullenix (eds.) *Talker Variability in Speech Processing*. San Diego: Academic Press, 145-165.

[3] Bybee, J. 2006. From usage to grammar: The mind's response to repetition. LSA Presidential address.

[4] Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psych. Review* 195: 251-279.

[5] Lacerda, F. 1995. The perceptual-magnet effect: An emergent consequence of exemplar-based phonetic memory. In *Proceedings of the 13th International Congress of Phonetic Sciences*, vol. 2, pp. 140-147.

[6] Stevens, K. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *J. Acoust. Soc. Am.* 111: 1872-1891.

[7] Traunmüller, H. 1981. Perceptual dimension of openness in vowels. *J. Acoust. Soc. Am.* 69: 1465-1475.

[8] Miller, J.L., Grosjean, F. 1981. How the components of speaking rate influence perception of phonetic segments. *J. Exp. Psy: Hum. Perc. and Perf.* 7 (1): 208-215.

[9] The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus (TIMIT). http:/www.ldc.upenn/readme_files/timit.readme.html

[10] Gütig, R., Sompolinsky, H. 2006. The tempotron: a neuron that learns spike timing-based decisions. *Nature Neuroscience* 9 (3): 420-428.

[11] Boersma, P, Weenink, D. 2007. Praat: doing phonetics by computer (Version 4.5.16) [computer program]. Retrieved February 18, 2007, from http://www.praat.org/

[12] Hawkins, S. 2003. Roles and representations of systematic fine detail in speech understanding. *J. Phonetics* 31: 373-405.