Speaking rate effects in a landmark-based phonetic exemplar model

Travis Wade, *Bernd Möbius*

Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany

{travis.wade|bernd.moebius}@ims.uni-stuttgart.de

Abstract

In this study we describe a model of speech perception in which neither speaking rate nor lower level temporal cues are considered explicitly. Instead, newly encountered speech signals are encoded as sequences of detailed acoustic events specified in real time at salient landmarks and compared directly with previously heard patterns. When presented with obstruent-vowel sequences occurring in the TIMIT database, the model performs similarly to humans in relying on temporal information for consonant and vowel recognition—and interpreting this information in a rate-dependent manner—when non-temporal cues are ambiguous; and by being adversely affected by local rate variability. These results indicate that compensation for speaking rate in human perception may follow implicitly from even modest knowledge of the robust correlations between temporal and other properties of individual speech events and those of their surrounding contexts, and do not require special normalization processes.

1. Introduction

Changes in speaking rate are a major source of variability in speech production. In instances where linguistic information is cued by temporal properties of the speech signal (consonant voicing, gemination, vowel quantity, etc.), this variability is potentially problematic, since it often affects these cues such that they lack invariant first-order descriptions. Fortunately, though, human listeners tend to compensate for these effects by perceiving temporal properties of speech in a rate-dependent manner. Quite a bit of research has attempted to characterize the normalization processes that would appear to drive this compensation, in terms of, e.g., under what conditions they operate, whether they result from speech-specific or general auditory mechanisms, and how speaking rate may be monitored and encoded [e.g. 1-4]. In this study, we test a very different possibility, namely, that explicit rate normalization does not occur, nor is speaking rate directly tracked or considered at all. Adopting a strong exemplar-theoretic position, we hypothesize instead that temporal (and other) properties of newly perceived sounds are directly compared with those of previously heard sounds, and that speech understanding is based on this comparison without any further abstraction. Critically, such a comparison must take into account surrounding context information, so that patterns of covariation between (temporal and other) cues in context and those directly related to a given category or contrast are preserved during recognition.

It is almost universally assumed that some type of perceptual rate normalization is necessary, since it would be impossible to represent *every* speaking rate in memory. Phonetic exemplar models provide a basis for testing this assumption empirically, since they assume that categorizations are based on finite distributions of memories that would not provide a continuous representation of (for example) speaking rate. Several phonetic exemplar models have been proposed [e.g. 5, 6] to show (among other things) how human listeners might retain detailed acoustic information while seemingly abstracting away from this information during linguistic processing. However, these proposals have been much less specific concerning the encoding of context information, especially where context is distributed over, or critically incorporates, the temporal dimension. Therefore, in the present study we present a new, context-oriented exemplar model. This model is described in detail in the next section. We then show that, based on realistic acoustic data, the model—and therefore the statistics of the language environment—implicitly predicts the following four facts about the perception of rate-varying speech: (1) when making phonetic decisions based on primarily temporal cues, listeners tend to interpret these cues relative to the surrounding context rate, in a manner that mirrors, and thus tends to compensate for, rate-related variability in speech production [1]; (2) when other cues (spectral, amplitude, etc.) are available in addition to temporal cues, the perceptual effects of context rate are attenuated and may disappear when the other cues are very naturalistic and unambiguous [2]; (3) the same temporal context information that affects perception of the temporal properties of one sound must simultaneously serve as (primary or indirect) evidence related to nearby sounds (e.g. the duration of a vowel may serve as context information for interpreting a preceding consonant but may also be required to signal the identity of the vowel itself); and finally (4) perception of sounds is affected by variability in speaking rate. Notably, increasing rate variability can negatively affect the accuracy of perception [5]. The first three effects relate mostly to apparent normalization processes; the fourth has often been discussed in relation to exemplar memory for sounds as well [e.g. 6], since it might reflect the increased cost of encoding this variable rate information in memory. We will suggest that a simpler mechanism is responsible for all four effects.

2. Model

Our model adopts a pure acoustic exemplar approach to representation and comparison. Perception does not involve segmentation or structural analysis of incoming speech, nor is any distinction made between linguistic and extra-linguistic aspects of the signal. Higher-order information such as speaker identity or speaking rate are not extracted or considered; in fact, temporal cues are not explicitly encoded at all. Instead, perception makes use of a memory containing an ordered collection of richly specified, real-time acoustic descriptions of previously perceived sounds, not unlike a continuous recording of previous auditory input. This memory signal is sparsely "annotated" with connections to records of other events that originally co-occurred with the acoustic pattern. Realistically, we imagine these events would represent varied, perhaps heterogeneous, non-hierarchical collections of visual, tactile, and motor occurrences in

addition to more abstract linguistic analyses or identifications. For the purposes of the present study, since the effects we model all involve phonetic categorization, we conveniently assume that such collections can be represented simply by traditionally described feature or segment labels. In the model, then, perception involves (1) comparing a newly encountered acoustic signal in space and time with the entire memory, and (2) identifying feature labels occurring near regions of maximum similarity.

Acoustic descriptions in our model take the form of potentially informative parameter values extracted at salient landmark locations in the signal [9]. This representation is selected less as a biologically plausible model than as a simple, transparent representation of some of the acoustic information that is likely to be of importance during perception, and the distribution of this information over time. In the current study we are interested in describing the perception of stop consonants in relation to neighboring vowels. Therefore, we considered three types of landmarks: abrupt consonantal (AC) landmarks occurring at consonant release bursts, abrupt (A) landmarks occurring at the onset of voicing, and vowel (V) landmarks occurring at points of minimal vocal tract constriction.

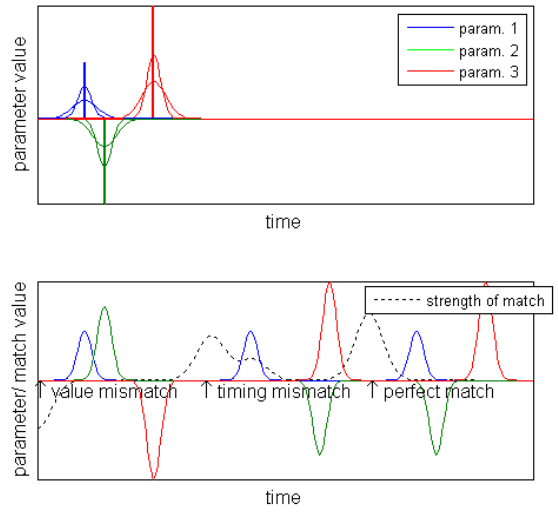

Figure 1: *Encoding, forgetting, comparing landmark-based descriptions. Top: 3 arbitrary parameters from an AC-A-V sequence before encoding (vertical lines), after 1 and 3 forgetting cycles (smoothed lines). Bottom: 3 signals in succession, with summed cross-correlation with the original signal. Arrows designate locations of labels; sequences match to the extent that landmarks with similar parameter values occur at similar intervals.*

In the model, temporal information is encoded in the locations of acoustic landmarks with respect to each other in time; spectral, amplitude and other information define the shape of the memory signal at these locations, in the following manner. The signal contains as many dimensions as there are parameters to be considered. Each dimension takes the (normalized) value of the relevant parameter at the location of the landmark where it is measured, and is zero elsewhere. Thus, memory resembles a multidimensional point process, with values in different dimensions occurring at different times, depending on the temporal distribution of landmarks. Examples of such signals are shown in Figure 1. Neither the acoustic parameter value nor the location of the corresponding landmark is encoded or remembered perfectly. Imperfection and the forgetting process are modeled by a single mechanism: the entire memory is smoothed with a narrow Gaussian window, such that values tend to regress slightly toward zero and their locations become distributed over time. During perception, a similarly encoded stimulus is compared with the entire memory by cross-correlation between the two signals, and connections to feature labels are activated as a function of their proximity to peaks in the resulting similarity function. Newly encountered sounds result in correlation peaks to the extent that the measured parameter values, and the relative locations of the landmarks at which they are measured, are similar. Again, however, no explicit temporal measurements are made.

3. Simulation

3.1. Acoustic data and training stimuli

The model was trained with acoustic data extracted from CV sequences in the training portion of the TIMIT database. Specifically, we included all of the sequences of stop consonants followed by the vowels [ɛ] and [æ], across syllabic/prosodic status, gender and dialect, a total of 1453 tokens. (Only [s]-voiceless stop clusters were omitted, since the de-aspiration that occurs in these cases is considerable and predictable by context but was not the focus of the current study.) 12 acoustic parameters were extracted and encoded in memory: at the AC landmark (taken to be the beginning of the consonant burst segment labeled in the database), the first three moments of the DFT distribution and the derivative of intensity (the difference between its value at the burst and 80 ms later) were included. At A landmarks (taken to be the beginning of the labeled vowel segment), the derivatives of f0 and the first three formants (again using an 80 ms difference) were included. Parameters considered at V landmarks (taken to be the maximum in the envelope of the CV below 500 Hz) were f0 and F1-3 values. F0 was estimated using an autocorrelation-based algorithm and formants using an LPC-based method, both as implemented in [10]. Spectra and intensity values were based on relatively long 50 ms Hamming windowed segments, to conservatively model sensitivity to change in these parameters over time. All frequency measurements used a linear Hertz scale.

Since there is no appreciable rate variability in the TIMIT corpus (or any similar, suitably labeled database that we are aware of), this variability was introduced artificially, employing linear expansion/compression of landmark offsets. All stimuli were presented to the model at three rates, by multiplying the distance from AC to A and V landmarks by 0.5, 1, or 2. While this is certainly not a realistic representation of rate differences (e.g. it is well known that different segments are affected differently by changes in rate), it sufficed to simulate variability without introducing any artifacts that a more data-oriented method might have.

Figure 1 shows the resulting distributions of landmark-to-landmark offsets, and thus the dependence of VOT on speaking rate, in the training set. While it was not the focus of the study, the acoustic measurements themselves contained substantial cues to both consonant and vowel identity. Voiced and voiceless consonants differed significantly ($p<0.001$) in terms of all six parameters measured, and [ɛ] and [æ] differed in all three formant locations.

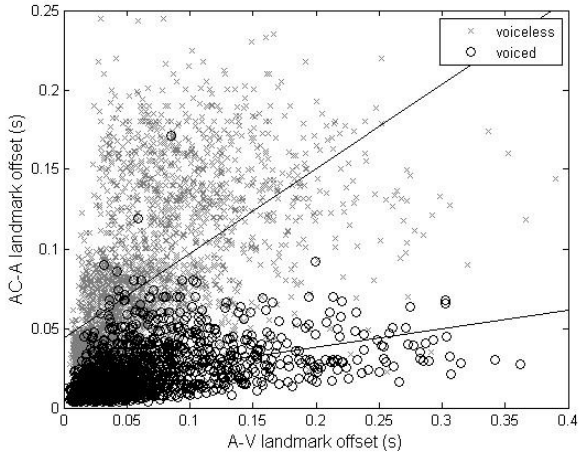

Figure 2: *Distribution of training data by landmark offset.*

These data were used to create a 12-dimensional memory signal, as described in section 2. In each dimension, values were first normalized to a mean of zero, standard deviation one. For each dimension for each CV sequence, a signal was created that consisted of zeros everywhere except at the relevant landmark location, where it took the normalized parameter value. These sequences were sampled at 300 Hz and padded with 100 ms of silence. The resulting signal was then convolved with a Gaussian window, s.d. 5 ms, normalized to a total power of 1.0, and concatenated to the end of the memory sequence. Consonant voicing and vowel height values (1 or 0) were marked at the beginning of a CV.

3.2. Test stimuli and procedure

Perception consisted of locating the k(=9) peaks in the cross-correlation function of an analogously constructed test stimulus with the entire memory sequence, and averaging the consonant voicing and vowel height values labeled nearest the corresponding point in the memory signal.

A single set of test stimuli was used to simulate effects (1)-(3) above. 800 items were selected at random from the same set of acoustic measurements used to create the memory. Each value set was then used to create 10 test stimuli: the AC-to-A landmark offset (signifying VOT) was varied from 40 to 200 ms in 5 steps, and the AC-to-V offset (indirectly signifying syllable length or speaking rate) was either 100 or 300 ms. Effects (1) and (3) were observed by comparing activated [voiced] and [low] features as a function of the AC-V offset. Effect (2) was simulated by measuring the magnitude of this same rate effect as a function of the ambiguity of the parameter values represented at consonant landmarks, as follows. Each token was rated as ambiguous for voicing in terms of parameter values (i.e. disregarding temporal patterns) by comparing the squared distance of these values from the center of the distribution of voiceless sounds with the distance from the center of the voiced distribution. The difference between values was then taken as a measure of "voicedness", with large positive values representing sounds that were much more similar to voiced than voiceless values, and negative values corresponding to more [-voiced] sounds. The absolute value of the voicedness measure was used to denote ambiguity: larger values indicate that the token is much more typical of one category than another, and thus less ambiguous. This value was then compared with the difference in [+voiced] activations generated by the resulting stimulus set as a function of the AC-V offset, averaging over the AC-A continuum.

To simulate effect (4), a similar test set was used. However, since the rate variability effect depends on relations between individual test set items, it was necessary to include test items in the memory as well. 500 items were selected randomly from the acoustic parameter set, and each item was used to generate one fast (AC-A offset 0ms, AC-V 100 ms) and one slow (100, 200 ms) token. In three conditions, either a fast, slow, or a randomly selected token was presented to the model, as described above. Following perception, this item was appended to the end of the memory sequence, which then underwent one forgetting cycle (convolution with a Gaussian window of s.d. 2.5 ms). To simulate an inverse correlate of response accuracy, we used the average of the absolute correlation values corresponding to the nine largest peaks in the cross-correlation function as a measure of absolute activation of the selected exemplars.

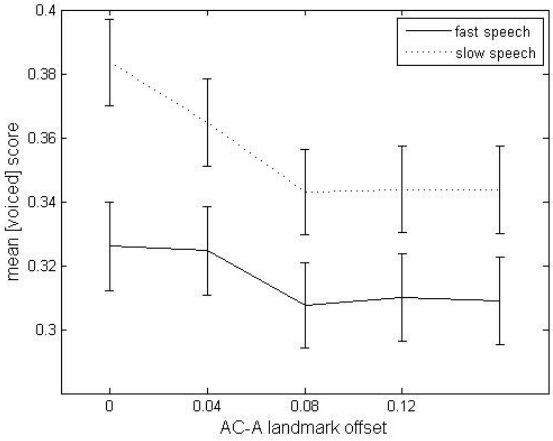

Figure 3: *Raw averages of activated [voiced] labels depending on AC-A landmark offset (VOT) and AC-V offset (speaking rate)*

4. Results and discussion

4.1. Offset effects on consonant categorization

Figure 3 shows the model's perception of consonants based on AC-A landmark offset (VOT) and AC-V offset (speaking rate). Raw averages of activated [voiced] values are shown. These results cannot be taken to represent phonetic categorization directly, since they do not take into account (1) the prior probabilities of voiceless and voiced sounds in the database (voiceless sounds were about twice as frequent), or (2) the non-linear relationship between categorization probability and perceptual similarity. Nevertheless, they demonstrate the essential points that would hold in an appropriate transformation: general reliance on VOT at both rates, and an effect of speaking rate. Like human observers, the model considers tokens as more voiceless for a given VOT at faster speaking rates. The mechanism for this is straightforward: shorter AC-A offsets were likely to be more similar to [-voiced]-labeled memory tokens that also had short AC-V differences, because these two offset values covaried in memory. No normalization or overt specification of rate (or even VOT) is required.

4.2. Non-temporal cues and rate effect magnitude

While consonant voicing perception was related to the AC-A offset, it also strongly correlated with the parameter voicedness measure described above ($r=0.87$; $p<0.001$), indicating that the spectral and amplitude parameters we included were also reliable cues to consonant voicing. Figure 4 shows the size of the rate effect as a function of ambiguity. As seen in the figure, the robust rate effect seen for the more ambiguous stimuli and reflected in figure 3 nearly disappears for stimuli with unambiguous spectral and amplitude cues. This is also in line with data on human observers [2];

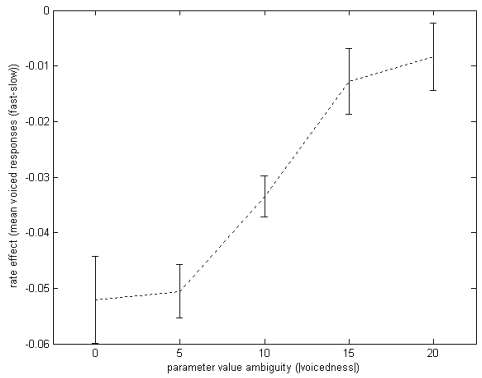

Figure 4: *the size of the rate effect (difference in mean [+voiced] responses across rate conditions) as a function of ambiguity of acoustic parameters.*

4.3. Offset effects on vowel categorization

Vowel classification was also influenced by AC-V offset: on average, vowels received higher [+low] scores with slower offsets ($p<0.001$). This is consistent with the acoustic patterns observed in the database and generally, and demonstrates that the same information that serves as rate context for one sound can simultaneously provide primary evidence for another sound.

4.4. Rate variability and activation strength

Figure 5 shows average activation strength across presentation conditions over the course of testing. After the first few trials, activation becomes lower (signifying lower accuracy) in the mixed rate condition. The mechanism for this is related to, but simpler than, the suggestion [e.g. 6] that exemplar-style encoding of rate information is a resource-demanding process. The forgetting process in the model dictates that, when there are very recently perceived items that are very similar to a stimulus along some acoustic dimensions (in this case timing properties of landmarks), these items have the potential to dominate the comparison process. The more similar recent items there are, the more likely a large positive (or negative) correlation value becomes, resulting in a more efficient recognition process.

5. Conclusions

We have presented a new phonetic exemplar model of perception in context, using it to demonstrate that several important facts about human perceptual compensation for speaking rate variability are straightforwardly predicted assuming detailed memory of a relatively small, highly variable set of productions. In reality, we assume that such a mechanism would represent only part of a massively hybrid system involving other modalities, processes, levels of representation. The model is therefore incomplete in that it does not incorporate these connections but depends on them, e.g. for the presence of feature labels. We are currently working to determine how these interfaces might be best characterized empirically, as well as to incorporate larger sets of acoustic parameters and more complicated landmark configurations into the model.

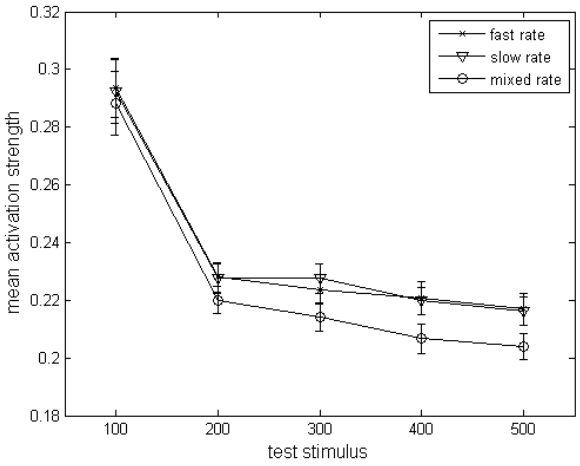

Figure 5: activation of contributing exemplars at different levels of rate variability during perception.

6. References

1. Summerfield, Q. "Articulatory rate and perceptual constancy in phonetic perception." J. Exp. Psych.: Hum. Perc. Perf., 7, 1981,1074-1095.
2. Shinn, P., Blumstein, S. E., & Jongman, A. "Limitations of context conditioned effects in the perception of [b] and [w]." Perc. Psychophys., 1985, 38, 397-407.
3. Wade, T., Holt, L. "Perceptual effects of non-speech rate on temporal properties of speech categories." Perc. & Psychophys., 2005, 67, 939-950.
4. Grossberg, S. "Resonant neural dynamics of speech perception." J. Phonetics, 2003, 31, 423-445.
5. Sommers, M., "Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude." J. Acoust. Soc. Amer., 96, 1994, 1314-1324.
6. Bradlow, A., Nygaard, L., Pisoni, D. "Effects of talker, rate, and amplitude variation on recognition memory for spoken words." Perc. Psychophys., 1999, 61, 206-219.
7. Pierrehumbert, J. "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, 2001
8. Johnson, K. "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, 1997.
9. Stevens, K. "Toward a model for lexical access based on acoustic landmarks and distinctive features." J. Acoust. Soc. Amer., 2002, 111, 1872-1891
10. Boersma, P, Weenink, D. Praat: doing phonetics by computer (Version 4.5.16), 2007, http://www.praat.org/