Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/phonetics

Syllable frequency effects in a context-sensitive segment production model

Travis Wade [*], Grzegorz Dogil, Hinrich Schütze, Michael Walsh, Bernd Möbius

Institute for Natural Language Processing, Universität Stuttgart, Azenbergstr. 12, 70174 Stuttgart, Germany

ARTICLE INFO

Article history:
Received 26 December 2007
Received in revised form
29 October 2009
Accepted 30 October 2009

ABSTRACT

In this study we describe a new model of how phonetic knowledge guides speech production. In the Context Sequence model, target acoustic patterns are determined based on selection of previously heard or produced sounds from a memory store. Since signals in the memory correspond to long stretches of continuous speech, individual speech sounds always appear in a larger context. A key property of the model is that the selection of exemplars for production is weighted by the similarity of the contexts in which they originally occurred to the current production context. In two simulations based on realistic amplitude envelope data extracted from a large single-speaker production corpus, we demonstrate that (1) optimal selection of context-appropriate segment-level exemplars requires consideration of about 0.5 s of context material preceding and following exemplars and (2) context-dependent production at this low level may be responsible for a range of frequency effects that have previously been assumed to involve word, syllable, and other higher levels of organization.

1. Introduction

Exemplar-based memory models provide a formal means of testing the notion that linguistic categories can be effectively described based directly on patterns in their occurrence during natural language use, and without reference to the abstract rules, processes and structures that have traditionally been invoked to explain these patterns. In such models, linguistic categories (such as words, phonemes, or parts of speech) are typically represented by collections of actual previously encountered instances of the categories, specified to an arbitrary level of—often redundant—detail. Identification and production, then, are driven by direct comparison of items within and between these collections, so that regularities in the detailed surface forms of category exemplars are preserved without being explicitly specified. Simultaneous simulation of more discrete linguistic phenomena (like phonetic neutralization) and more gradient ones (like speaker- or word frequency-dependent sub-phonemic acoustic differences) based on exemplar representation have been taken to suggest that "usage-based" accounts provide an accurate, parsimonious description of linguistic competence and performance (Bod, 2006; Bybee, 2002, 2006; Foulkes & Docherty, 2005; Hawkins, 2003; Pierrehumbert, 2001; Port, 2007).

Exemplar approaches have shown success at accounting for several key observations in language production, perception, memory, acquisition and historical development (Goldinger, 1997; Hay, Nolan, & Drager, 2003; Johnson, 1997, 2006; Lacerda, 1995; Lively, Logan, & Pisoni, 1993; Pierrehumbert, 2001). A serious limitation of most specific models formalized so far, though, is that they have generally treated categories at various levels of organization as separate, independent events and have generally not considered the fact that speech unfolds in the temporal dimension, or the effects that temporal context information has on the specification of individual events. At a phonetic level of description this is especially limiting, since the speech signal is characterized by essentially ubiquitous change in the spectrum over time. If exemplar theory is to provide a comprehensive model of phonetic knowledge that can be used to study production and perception of connected speech, then exemplar comparison and selection should reflect this fact in a meaningful way. In this paper, we demonstrate and quantify the importance of surrounding context information in specifying segment-level exemplars, and show that this importance is modulated by the frequency of occurrence of the context, making estimations based on a large single-speaker corpus. We also describe a new exemplar-based perception/production model, the Context Sequence model, that reflects the temporal and the context-dependent nature of speech, and demonstrate how consideration of context at a local level automatically predicts aspects of production that have previously been attributed to more abstract, more complex hierarchical organization. First, we motivate and introduce key properties of the model by reviewing some related findings from speech perception.

1.1. *Context in exemplar-based perception and implicit normalization*

A strength of exemplar accounts of human speech recognition is how they deal with the inherently multidimensional nature of

[*] Corresponding author. Tel.: +49 415 394 3100.
E-mail address: travis.wade@ims.uni-stuttgart.de (T. Wade).

0095-4470/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2009.10.004

(proposed) phonetic units such as segments or features. Since potential cues in the speech signal show wide and complex patterns of variability across different productions, speakers, etc., abstract categories have notoriously defied low-order acoustic or perceptual descriptions. This has led researchers over the years to posit complex normalization processes in mapping from acoustic to linguistic representations in order to account for humans' apparently effortless recognition. Such normalization is implicitly predicted by exemplar approaches, where the "overspecification" of actual productions in memory results in the preservation of predictable patterns during decoding or further production (Johnson, 1997, 2006; Pisoni, 1997). That is, the use of one type of acoustic cue (such as a formant frequency) in interpreting—or planning the production of—a sound need not be explicitly modulated depending on other, co-occurring cues (fundamental frequency or voice quality, for example), because an effective pattern classification mechanism will take into account the covariance structure with which the different cues are represented in the language environment. Thus, consideration of information that would normally be considered irrelevant, or even non-linguistic, actually provides dimensionality that is necessary for proper interpretation of phonetic events.

Conspicuously absent in exemplar models developed so far, however, is much reference to the acoustic contexts in which phonetic events occur. During perception, dimensions along which a segment, for example, must be considered for successful classification include not only potential cues that occur during the production of that segment, but also those in the temporal regions adjacent to it, as well as the temporal organization of the signal itself (e.g. the speaking rate). Production and perception of segments are both thoroughly context-dependent due to the limited types and rate of change in the configuration of the vocal tract that may occur over time and the sensitivity of the auditory system to contrast and dynamic patterns rather than to absolute measures (Kluender, Coady, & Kiefte, 2003; Lindblom, 1963; Lindblom & Studdert-Kennedy, 1967). In fact, it is probably fair to say that very little phonetic information at a segmental time scale is very meaningful outside the context in which it originally occurred. This problem does not disappear if larger units (such as words) are taken to be the objects of exemplar-style representation, since context exerts influences over an extended period of time (Ladefoged & Broadbent, 1957; Summerfield, 1980; Holt, 2005).

We recently suggested that these facts are consistent with an exemplar approach to speech perception, but that they imply that context and temporal order must play a role in exemplar specification and comparison. In a series of simulations (Wade, 2007; Wade & Möbius, 2007), we proposed that exemplars consist of sequences that represent stretches of speech much longer than the units of interest, perhaps corresponding to entire utterances. These sequences preserve both local, detailed spectral and amplitude information and the temporal organization of this information.[1] Perception, then, involves comparing newly encountered patterns, along with their surrounding context material, to sequences of similar length in memory. From one perspective this is an extreme simplification of the types of cues involved, since dynamic patterns are represented only indirectly in the temporal adjacency of similar (or critically differing) configurations. In this way, though, correlations between different types of spectral information in adjacent regions (like phonetic context effects related to coarticulation), between different types of spectral information at the same location (like speaker or gender effects), and between spectral information and temporal organization (like speaking rate effects) are all potentially accounted for in the recognition process. Simulating an exemplar memory based on sequences from a large, multi-talker speech corpus, we were able to model several key aspects of humans' perceptual dependence on context, without reference to explicit normalization processes.

1.2. *Context in production and frequency effects*

Since the set of sounds that is produced involves (for the most part) the same patterns of covariance and redundancy over time that characterize the set of sounds that are perceived, speech production must similarly take context into account. Exemplar theory addresses the perception–production link in an intuitive way, assuming that production involves a selection from the same store of acoustic memories that is used during perception (Bybee, 2002, 2006; Goldinger, 1997; Pierrehumbert, 2001). Based on our perception simulation results, we propose that this memory store is effectively composed of representations of fairly long stretches of speech. Individual "exemplars"—that is, portions of such a representation that might be associated with a category label (for example, segment, feature, or word)—are stored adjacent, or closely linked, to the contexts in which they originally occurred, including neighboring, overlapping, or encompassing exemplars. Selection of exemplars for production is informed by this context information. Just as exemplar perception models identify newly encountered sounds by comparing them with stored members of existing categories, the basic assumption of the Context Sequence account is that selection of a stored category exemplar for production is weighted by the similarity of the exemplar's original context with the relevant neighboring sounds in the current production context.

Such a context-dependent selection process might have far-reaching effects. The simulations described in this paper relate to two classes of observations concerning the frequency of occurrence of different categories in a language. It is often reported that very frequent syllables, words, and collocations seem to take on an "autonomous" status with respect to acoustic variability and historical change, involving type-specific patterns of variability and a resistance to generalization. For example, vowels and some consonants are more likely to be weakened or deleted in more frequent words and phrases (Bybee, 2001, 2002; Hooper, 1976) than in less frequent contexts. Perhaps relatedly, more frequent words and syllables are consistently produced faster in speeded tasks than less frequent ones (Cholin, Levelt, & Schiller, 2006; Jescheniak & Levelt, 1994; Levelt & Wheeldon, 1994).

Current exemplar-based production models account for the historical acoustic patterns almost trivially, since they assume that acoustic mutations are introduced at the level of the individual production, with lenition and other forms of variability compounding over the course of large numbers of productions (Bybee, 2001, 2002). The syllable latency results have often been taken (see also Levelt, Roelofs, & Meyer, 1999) to suggest that complete articulatory gestural programs are memorized for the most frequent syllables but computed online from lower-level specifications for less frequent ones, requiring more processing and, presumably, less efficient production. We recently proposed that a closely related multi-level exemplar-based process may be at work and demonstrated that such a process could additionally account for observed patterns in syllable length variability. Schweitzer and Möbius (2004) reported that, in a large single-speaker production corpus originally designed for unit-selection speech synthesis, relative syllable length was well predicted by

[1] In these previous simulations, local information included the values of cues such as formant frequencies at salient landmark locations in the signal (Stevens, 2002) and temporal information amounted to an encoding of the temporal locations of the landmarks at which measurements were made.

the summed relative lengths of constituent segments for the most infrequent segments, but much less so for the most frequent segments. As part of a model designed to account for both syntactic and phonetic exemplar-based production, Walsh, Schütze, Möbius, and Schweitzer (2007) and Walsh, Schütze, Wade, and Möbius (2007) suggested that selection always involves a competition between units at neighboring levels of an organizational hierarchy. For example, during speech production an attempt is first made to select syllable-level exemplars; if the available data at this level are too sparse, production "drops down" to proceed at the level of the individual segment instead. Thus, in corpus-based speech synthesis, it is possible to preserve the forms of entire syllables and words only for the most frequent units, since the number of infrequent units would require impossibly large data sets. In much the same way, the human production system might rely on complete unit-level information where possible and resort to more local, constituent-level specification where necessary.

Since different, generally larger-scale patterns of variability are likely to result from higher-level selection, a multi-level selection process might provide additional explanation for the frequency-specific patterns of variability discussed above (Bybee, 2001, 2002; Hooper, 1976). Although accounts of this sort point to variability and evolution across productions of entire syllable-level (or word-level, or collocation-level) units, it might be noted that it is generally the constituent articulations composing these frequent sequences that actually differ when compared with their counterparts in less frequent sequences. Thus, a slightly different characterization of the findings is that the variability of these constituents (segments, for example) is conditioned by the frequency of the acoustic contexts in which they occur. This is a nontrivial distinction with respect to the explanatory power of the exemplar selection process. Previous explanations assume an explicit, hierarchical link between two or more levels of linguistic structure, and an explicit transfer of information (such as word frequency or amount of lenition) from a higher to a lower level. While we will not suggest that higher-level (or greater-length) units do not have a place in exemplar-based production, this particular explanation seems to import a lot of structure from traditional linguistic theory, some of which may be unnecessary if we already assume—based on empirical data—that exemplar selection is influenced by (acoustic) context.

Under the right circumstances, a context-based model working exclusively at a local (e.g. segment) level also predicts that more frequent sequences (such as syllables) are produced more efficiently and with sequence-specific patterns of variability or lenition. Where the surrounding context of the segment currently being produced is very similar to many sequences in memory—for example, in the middle of a frequent syllable—it should be quick and easy to find a segment-level exemplar that is appropriate for this context. And since this exemplar will likely have been originally produced as part of the same syllable that is now being generated, any previous lenition or other variation will be preserved, and possibly compounded on, in the new production. Where the syllable context is rarer, it is less informative as to which segment exemplars are most appropriate (since it is equally dissimilar to most sequences in memory), so selection will be less constrained and less efficient, with the eventual output probably tending more toward the center of the overall segment distribution. Such frequency-based differences might be thought of (at least in this context) as the emergence of a competition across hierarchical levels.

In the next sections, we describe the Context Sequence model of exemplar selection in more detail, and briefly discuss properties of the language environment that would drive the frequency effects we have just predicted. We then describe two simulations, based on productions from a large single-speaker corpus, demonstrating quantitatively that (Experiment 1) these patterns do occur in language and (Experiment 2) they do result in the emergence of syllable-level exemplar selection if the proposed model has access only to more local, segment-level specification and context.

2. A context-based production model

Following previous exemplar accounts (esp. Pierrehumbert, 2001), we view the production process simply as a selection from a set of remembered tokens, followed by a probabilistic degradation process. As described below, the degradation mechanism implemented here is somewhat abstract and could be thought of as representing either lenition or adherence to a historical trend in production. In reality, a combination of many such processes, as well as more random sources of variation, probably contribute to produced forms.

In the present study, we model the production of entire utterances, and not just the selection of unspecified or high-level values related to isolated events. We propose that there is an iterative process of specification in context, whereby acoustic information is filled in bit by bit (and generally left-to-right), based on the evolving production context. In the simulations presented here, specification and production are modeled exclusively at the segmental "level". This choice was motivated more by practical than theoretical concerns. We do not actually assume that segments or phonemes comprise the primary level (or even one fully independent level) of selection during human speech production; working at this level, though, provides for the selection of acoustic information that is—relative to syllables, words, and phrases—local and highly context-dependent. Of considerable practical importance for replicating and extending our results, the segment level is normally the most fine-grained level of labeling found in speech corpora. In Section 5.1 we return to the question of what effect this choice has on the implications of our findings, and what a more realistic level of selection might be.

Regardless of position in a word or phrase, production always progresses at the segment level. For each successive segment in an utterance, a cloud of segment exemplars is generated and each token is weighted based on the match between the current production context and the context in which it was originally produced. An exemplar is selected based on this weighting, degraded slightly, and produced. Its acoustic properties then become part of the current production context, on which selection of following segments is based.

2.1. The context-match comparison

Since both preceding and following context seem to be important in characterizing segments (Mann & Repp, 1980; Miller & Liberman, 1979; ʺOhman, 1965; Whalen, 1990), appropriate exemplar selection must take into account context on both sides of the segment currently being produced. The description above only straightforwardly accounts for left context (the acoustic specifications of recently produced segments). Acoustic information to the right of the current context would have to involve an estimation of what is likely to be produced in the future. While such an estimation might well occur, for the purposes of the current simulations we simplify the situation somewhat, and assume that two basic types of context information are available: acoustic information (left context, of the type discussed above) and what we will simply refer to as "linguistic" information, which encodes the identities (category names) of the segments

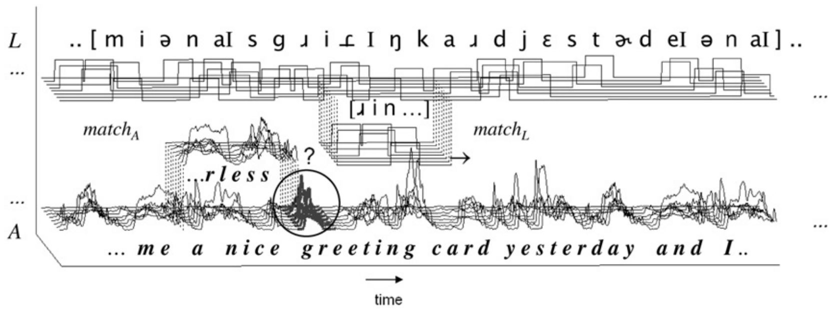
L
..[m i ə n aɪ s g ɹ i ɾ ɪ ŋ k a ɹ d j ɛ s t ɚ d eɪ ə n aɪ]..
...
[ɹ i n ...]
match_A
match_L
?
...rless
...
A
... me a nice greeting card yesterday and I..
...
...
time

Fig. 1. Determination of the context-match value by which a single segment exemplar is weighted for selection during speech production. The appropriate shape of the [g] in the phrase "colorless green ideas" currently being produced (center) is being estimated by comparing the just-produced acoustic context match$_A$ and planned immediately following linguistic context match$_L$ with the acoustic information A preceding, and linguistic information L following one exemplar of [g] (circled) in memory.

that will be produced in the following context. The exemplar weighting comparison, then, involves two parts. The acoustic information that originally preceded an exemplar is compared (match$_A$) with the current acoustic context (the sequence that has just been produced), and the names of the segments that followed the exemplar are compared (match$_L$) with the next planned segments in the current context. The two resulting scores are combined to provide an overall context match value:

$$\mathrm{c} - \mathrm{match} = \mathrm{match}_A(\mathrm{target} - m \ldots \mathrm{target}) + \mathrm{match}_L(\mathrm{target} \ldots \mathrm{target} + n)$$

where m and n represent the amount of preceding acoustic information and following linguistic information to be considered. Fig. 1 schematizes this process for the evaluation of a single existing exemplar of the category [g], which is currently being produced in the context of the (perhaps never before uttered) phrase "colorless green ideas". match$_A$ is calculated by comparing the just-produced acoustic information to the part of the signal that originally preceded the [g] being considered (represented in the figure by continuous values to indicate the detail with which it is likely to be encoded) and match$_L$ comes from comparing the next planned segments—represented as a binary function to indicate a more abstract nature (one possible implementation is described in Experiment 2)—to those that originally followed the [g]. This particular exemplar provides a fairly good context match, since the vowel-sibilant part of the acoustic sequence that originally immediately preceded it is similar to the vowel-sibilant sequence that has just been produced, and the [ɹi] sequence that originally followed it is identical to the next planned segments. Since exemplar selection is weighted based on the context-match score, this exemplar will be very likely to be chosen. More generally, tokens that are selected for production in a given context will tend to be the most appropriate ones—in terms of the statistics of the language, dialect, speaker, etc.—for that context. As a result, we hypothesize that configurations of segments relating to very frequent words, collocations, phrases, etc. may compound variability over repeated context-appropriate productions in the same way as if these higher levels had been involved explicitly.

2.2. Measuring the importance of context

If context-weighted selection is to model the speech production process in a realistic way, a necessary first question is how much context it is actually useful or appropriate to consider. Perception data suggest that acoustic information spread over several hundred milliseconds can play a role in determining segment identity (e.g. Holt & Wade, 2004). However, such results do not provide a quantitative estimate of how important the information at a given distance from the point of the current production context is in determining the appropriate output at that point. Experiment 1 represents an attempt to estimate this function for a limited set of contexts.

Fig. 2 demonstrates the intuition behind our approach to this estimation. Suppose that we select one target segment exemplar of a single phoneme category from a sequentially organized memory for speech like the one depicted in Fig. 1,[2] where the acoustic characteristics over each segment are represented simply as a vector of real values with expected mean zero. If we compare the target segment's acoustic characteristics with those of every other (analogously represented) segment in the memory—for example by taking the scalar or dot product (·) of each pair of vectors:

$$\vec{v} \cdot \vec{w} = \sum_{i\, v_i \cdot w_i}$$

We would generally expect that other segments of the same phoneme category as the target would show greater similarity with it, on average, than exemplars of other phoneme types. Further assuming (unrealistically) that the acoustic characteristics of segments of different phoneme types are randomly and independently distributed, the target's average similarity with exemplars of different phoneme categories would be zero. However, depending on factors including how the acoustic signal is actually encoded in memory, there would certainly be variability and probably substantial overlap in the distributions of similarity values seen between the target segment and other same-category and different-category segments.

A key assumption of the Context Sequence account is that a meaningful amount of this variability is predictable given the acoustic characteristics of *surrounding* segments, due to both local coarticulatory and longer-range context-specific production patterns. Suppose now that we take not only the target exemplar, but some of its surrounding context in addition, and compare the entire (now longer) resulting acoustic vector with the sequences of the same length centered around, again, all of the other segments in memory. The hypothetical distributions plotted in Fig. 2 demonstrate what might happen with contexts of different types and lengths. First of all, as the size of the context grows, there would be a general increase in the variance of total similarity values, since larger vectors are being multiplied and summed. This effect is visible in all panels of the figure, and completely accounts for the shape of the distribution of different-category similarity functions shown in the first panel (see figure caption for details); these sequences have zero average similarity to the target regardless of how much context considered. Target-category segments, on the other hand, have greater-than-zero

[2] More precisely, as A in Fig. 1, ignoring the L stream for the moment since we are only concerned with comparing existing examplars and not planning new ones in Experiment 1.

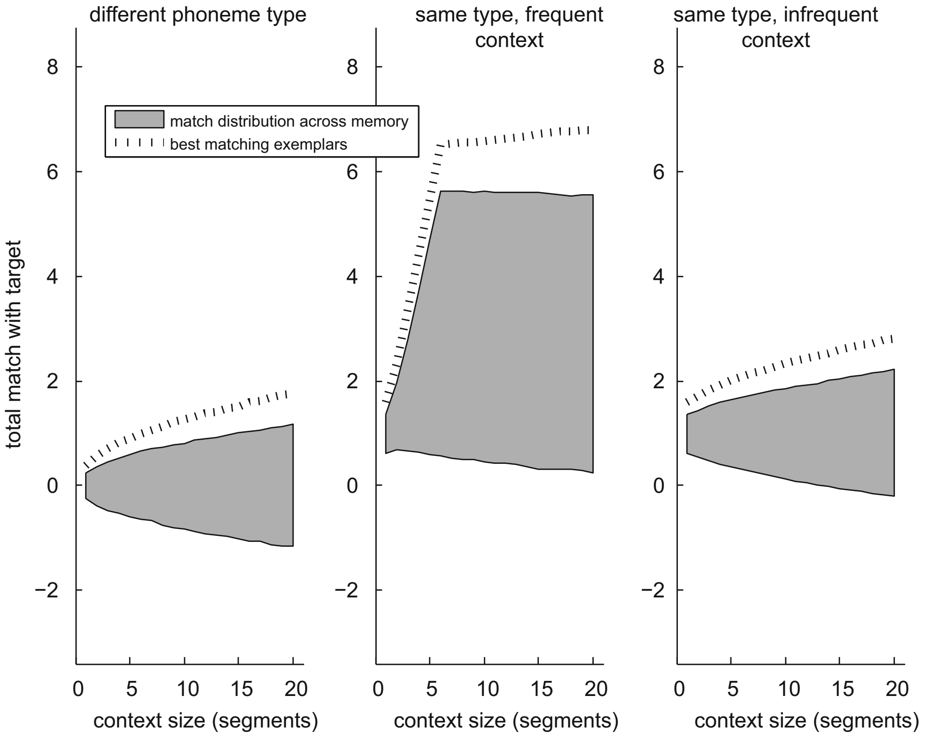

Fig. 2. Hypothetical distributions of similarity between arbitrary segment sequences in a temporally ordered acoustic memory with the sequence centered around a given target segment, depending on the segment category and the frequency of occurrence of its surrounding context as described in Section 2. The *x*-axis gives the size of the context in segments. For example, a context size of 4 corresponds to 2 segments before and 2 segments after the probe. The *y*-axis shows the cumulative similarity, the summed product of pairs of vectors with zero mean and variance. Different phoneme type comparisons (left panel) represent similarity between independently selected vectors, while for same phoneme type comparisons (right panels) the portions corresponding to the target segment were identical between the two vectors. For same-category vector pairs with frequent contexts (middle panel), values corresponding to an additional five segments adjacent to the target were held identical. The shaded areas represent expected 20th and 80th percentiles in the similarity score distribution; the dotted lines show the 90th percentile, representing matches that would likely dominate a selection process.

average similarity to the target segment. In addition, those target-category segments that originally occurred in contexts similar to that of the actual target will tend to have greater similarity still, and the similarity score for these exemplars will *increase* as context is added, since the context-to-context portions of the comparison will also be positive. Importantly, we would also expect differences in the distributions of target-category similarity values depending on the frequency of the context (the word, etc.) in which the target exemplar occurs. These differences can be seen in the right panels of Fig. 2. For very frequent contexts, the best similarity scores continue to increase as more surrounding material is added, since matches of entire words, phrases, etc. are likely. As the amount of context considered continues to increase past, say, the word or utterance level, the probability of a good match (a sequence of similar segments produced at a similar rate) eventually decreases even for the most frequent contexts, so that even the best similarity functions level off, continuing to become more variable but not higher on average. For less frequent contexts, while very local context information might enhance the match for the most similar exemplars, the match function should level off much earlier than for frequent contexts.

It is this difference that we predict may drive the frequency effects on lenition and production latency described above. It seems likely that the exemplar selection process is mostly driven by the best-matching contexts, since exemplars originally produced in inappropriate contexts are very unlikely to be selected unless there are no better matches. Therefore, it is the upper range of the distributions in Fig. 2 (also shown by the dotted lines in the figure) that are important in predicting frequency-dependent variability patterns. If the amount of context considered corresponds to about the center of the *x*-axis in the panels of Fig. 2, then selection of exemplars in frequent contexts will be informed by the match values, since exemplars with very high *c−match* scores will be encountered and dominate the selection process. As a result, another exemplar that originally occurred in the same context (in the same word, at the same rate, etc.) will probably be selected. For infrequent contexts, since very good context matches are much less likely, context-based weighting will have less influence, resulting in a flatter distribution of segment exemplars to choose from. In Experiment 1 below, we verify that these overall patterns actually occur, and attempt to associate them with actual time value estimates (the *x*-axis in Fig. 2), based on a corpus analysis.

2.3. Representation and comparison of acoustic content

Closely related to the lack of work on temporal and context issues in previous exemplar models, there has been little focus on a realistic representation of the speech signal in memory. Words have commonly been represented by constant-length random vectors (e.g. Goldinger, 1998), and phonemes and smaller units by one or a few isolated acoustic parameter values (formant frequencies, durations, etc.) that assume some sort of lower-level analysis or abstraction (Johnson, 1997, 2006; Pierrehumbert, 2001).

In previous simulations (Wade, 2007; Wade & Möbius, 2007), we have considered sequences of values of the latter type, representing changes in, for example, fundamental or formant frequencies over time. In the present study, we employ representations that more faithfully encode the speech signal as it unfolds over time without making specific assumptions about what types of cues might be extracted or which regions of the signal are the most important. Specifically, we consider the slowly varying amplitude envelopes of the signal across different frequency bands. There is much ongoing debate about the relative contributions of this type of information, and of the quickly varying temporal fine structure related to energy close to bands' center frequencies, to perception and memory for speech under different listening conditions (e.g. Lorenzi, Gilbert, Carn, Garnier, & Moore, 2006; Zeng et al., 2005). However, it is clear that (1) envelope signals represent information that is present, in various forms, throughout the auditory system and (2) they contain enough information to construct intelligible, fairly naturalistic speech when at least nominal spectral resolution (a few frequency bands) is provided for (Loizou, Dorman, & Tu, 1999; Shannon, Zeng, Kamath, Wyngonski, & Ekelid, 1995). Thus, in addition to being (relatively) compact and transparent representations, and lacking abstract dimensions that make assumptions about front-end analysis or estimation from the signal, envelope signals probably represent at least some of the information that is actually stored and considered by humans.

We also do not attempt to address in the model or analysis the question of how the temporal dimension is actually represented in neural terms, since this issue is currently far from being fully understood (e.g. Eagleman et al., 2005; Mauk & Buonomano, 2004, for recent reviews); we simply assume that signals are effectively stored as linear time sequences. Similarly, we model the exemplar comparison process simply as a cross-correlation between two signals. It is not clear how or whether such an operation is actually performed at a neural level, but since correlation processes are commonly invoked to account for a range of auditory phenomena including pitch perception and sound localization (Jeffress, 1948; Licklider, 1951), it seems reasonable to assume some approximation to this type of comparison.

3. Experiment 1

Experiment 1 estimates the importance of varying amounts of acoustic context in characterizing segment-level exemplars, by (1) selecting segments from very frequent and very infrequent contexts in a simulated memory, (2) computing the acoustic similarity of each of these segments with all of the other segments in the memory, and (3) measuring changes in the relative similarity of the best matching sequences as more surrounding context is included in the comparison. The output of this experiment will be used to specify a duration window for preceding context in the Context Sequence model that optimizes the selection of highly similar exemplars during production.

3.1. Acoustic data and memory composition

Acoustic data were envelope signals derived from a corpus of standard German sentences (Schweitzer, Braunschweiler, Klankert, Möbius, & Säuberlich, 2003; Schweitzer & Möbius, 2004), read by a professional speaker and annotated at segment, syllable, and word levels by forced alignment with manual checking. The corpus consisted of 2776 utterances sampled at 16 kHz. The envelope extraction process was typical of methods that are commonly observed to result in intelligible speech when the envelopes are used to modulate pure tone or band-passed noise signals (e.g. Loizou et al., 1999).[3] Utterances from the corpus (excluding preceding and following silence intervals) were first equated for root mean square amplitude, and then separated into eight logarithmically spaced frequency bands from 80 to 8000 Hz using 4th-order Butterworth filters. Amplitude envelopes for each band were estimated using the Hilbert transform, low-pass filtered at 50 Hz, and downsampled to 100 Hz for processing efficiency. The resulting dataset represented 12 180 s of continuous speech comprising 107 209 segments and 41 359 syllables, concatenated in arbitrary order to form a single, labeled eight-dimensional (corresponding to the eight frequency bands) "memory" sequence. Finally, in order that summed cross-correlation sequences between signals of different lengths would tend to center around zero (as in Fig. 2) and thus be more easily interpreted, the entire memory sequence was normalized by subtracting the mean in each dimension (band).

3.2. Procedure

One token each of the 50 most frequent syllables (average frequency 300, s.d. 264) and the 50 least frequent syllables (all frequency 1; taken arbitrarily from 1099 such syllables) in the corpus were first selected at random. From each of these 100 syllables one segment was then randomly selected. These segments represent the frequent- and infrequent-context target exemplars discussed in Section 2.2 and pictured in the right panels of Fig. 2. The segments themselves had approximately equal average type frequency regardless of syllable context (frequent context: mean segment frequency 5343, s.d. 2843; infrequent context: 4302, 3540). 2100 probe stimuli were then constructed by taking the acoustic material corresponding to these 100 segments, plus from 0 to 1000 ms of preceding and following context material, varied in 50 ms steps.

To estimate similarity scores for target segment-plus-context sequences analogous to those shown in Fig. 2, each probe was then compared acoustically with the entire memory sequence, as follows. First, a raw match function M the length of the memory was computed by cross-correlation of the probe envelope with that of the memory, summed across frequency ranges:

$$M(t_m) = \sum_{d=1}^{D} \sum_{t_p=1}^{T_p} probe_{d,t_p}\, memory_{d,t_p+t_m-1}$$

where t_m is the sample of the memory under consideration, t_p is the sample of the length-T_p probe, and d is the frequency range ($D=8$). Then a normalized match function M_n was created by adjusting M for overall amplitude in the memory sequence. This was to reflect the fact that speech events are better defined by spectrotemporal changes than by static patterns, and more practically to prevent regions of higher average magnitude from obscuring close matches (for example, to prevent a relatively weak consonantal sequence from correlating better with a long stretch of large positive values corresponding to a prominent stressed vowel than with a very similar version of the same consonant). This was accomplished by dividing M by a running estimate of local amplitude: a full-wave rectified version of the original memory sequence, summed across dimensions and then smoothed with a T_p−length rectangular window,

$$M_n(t_m) = \frac{M(t_m)T_p}{\sum_{d=1}^{D} \sum_{t=t_m-T_p/2}^{t_m+T_p/2} |memory_{d,t}|}$$

Finally, the maximum match value in the portion of the M_n that corresponded to each segment in the corpus, M_{nmax} was recorded as a measure of the similarity of the probe sequence with the sequence centered around that segment.

For each segment type in the corpus, the top 10% of the resulting maximum match values corresponding to tokens of that

[3] It was verified that sounds achieved by multiplying similarly band-passed noise signals by the envelopes used in the study were generally easily recognizable as the intended utterances.

segment were averaged to form a best-match score for the type. These values, calculated at different context lengths, are analogous to the "best-match" thick dotted lines shown in Fig. 2. Again, best-match values might be thought of as corresponding to the exemplars that would typically dominate a selection process during speech production or a matching process during recognition.

It is important to note that, since match (and best-match) scores always relate to the *maximum* correlation value from a sequence the length of the segment being compared, these values skew in the positive direction and are greater than zero on average. Since this would result in distributions different from those shown in Fig. 2 (where it was simply assumed that values would center around zero) and obscure the timing of any frequency-related differences in the inflections in target-segment match functions like those shown in the figure, one final transformation was made. A "same-type advantage score" for each probe sequence was estimated in order to quantify the usefulness of a given amount of context in characterizing the target segment's phoneme type while normalizing for the variability that would be expected for sequences of that size. This score was calculated by subtracting the average best-match score for non-target categories from the best-match score of the target category, and then dividing by the standard deviation of the non-target best-match scores.

3.3. Results and discussion

Fig. 3 shows a representative subset (100 randomly selected examples of each match type) of the match-to-context-size functions M_{nmax} that resulted from the analysis. Although, again, values are skewed in the positive direction and greater than zero on average, the figure seems to demonstrate the main features of the predictions shown in Fig. 2. At the lowest context sizes, the two same-type conditions show better average matches than the different-category condition. At about 300 ms, the two same-type distributions diverge, with the low-frequency context leveling off sharply and most of the high-frequency contexts continuing to increase up to about 0.5 s.

To quantify this effect, the same-type advantage described above is shown across context sizes in Fig. 4. Here it is clear that there is a same-type advantage (indicated by positive values in the figure) regardless of context frequency that increases for contexts of up to about 0.1 s. As in Figs. 2 and 3, this advantage then levels off for infrequent contexts but continues to increase up to about 0.5 s for frequent contexts.

Thus, preceding and following context of up to about half a second continues to provide additional information about the appropriateness of exemplars in frequent contexts, but only very locally for infrequent contexts. Perhaps interestingly, the short (0.1 s) sequences that characterize low-frequency contexts are on the order of the average segment length in the corpus, while the longer context sequences that characterize frequent contexts are closer to a typical syllable length. This means that, to emulate human performance, the context model should consider context somewhere between 0.1 and 0.5 s. The next section describes a more fully elaborated production model that makes use of this fact.

4. Experiment 2

Experiment 2 was designed to test whether frequency-of-context effects like those observed in Experiment 1 could actually drive historical acoustic changes and differences in production latency. A version of the Context Sequence model described in Section 2 was constructed to generate segment-level exemplars. We then observed its production of segments corresponding to a long string of *syllables* whose composition and frequency were determined by a grammar that was external to the production model (i.e. the model only had access to segment-level information). The exemplar memory on which production was based consisted of the segments that had been produced so far at a given point in the simulation. At the beginning of the simulation, segments were produced based on predefined, default sequences. As the exemplar memory grew, production came to rely on the segments that had previously been produced and added to the memory. Each production involved a subtle systematic distortion of the exemplar on which it was based, resulting in gradual

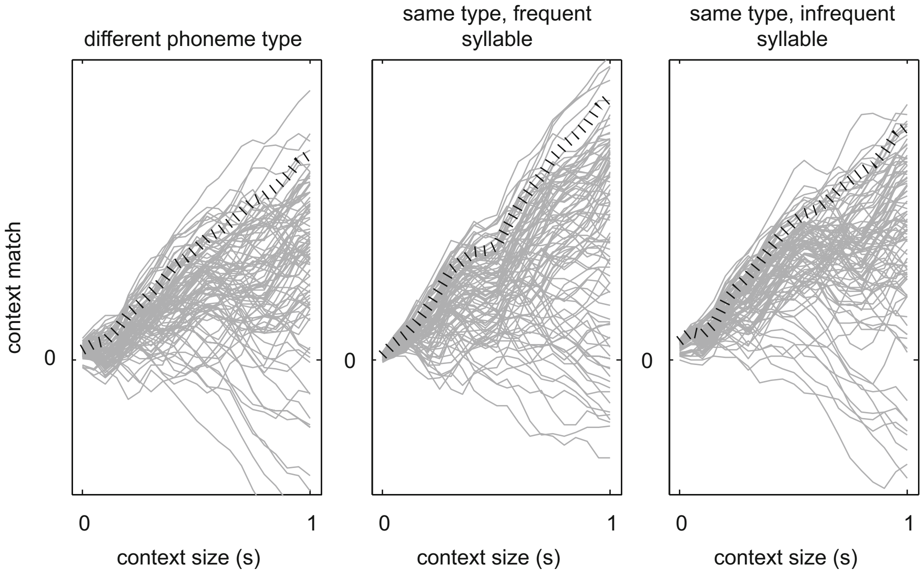

Fig. 3. Sample match functions from Experiment 1. The *x*-axis gives the size of the context in seconds. For example, a context size of 0.4 s corresponds to 0.2 s before and 0.2 s after the probe. The *y*-axis (scale is arbitrary since it depends on the scaling of the signals in memory) represents M_{nmax} scores (see Section 3.2). As in Fig. 2, the left panel represents comparison across phoneme types and the right panels show same-phoneme-type for (frequent- and infrequent-context) comparisons. Thin gray lines represent individual match curves for arbitrarily selected tokens, and the dotted line shows the mean of the top 10% of matches for each segment.

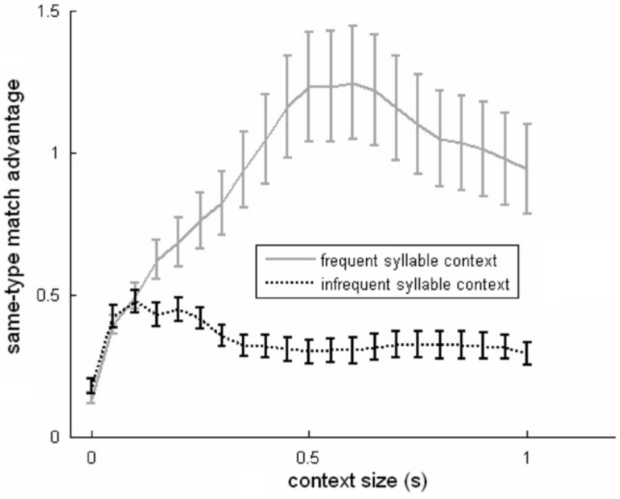

Fig. 4. The same-type match advantage (with standard errors reflecting the 50 items of each type considered) at different context sizes.

category change over time. Of interest was whether patterns of acoustic variability and exemplar selection emerged that were specific to the frequency of the contexts in which segments occurred. Our predictions were that, relative to (the same segments in) infrequent contexts, segments produced in frequent contexts would (1) be produced more quickly and efficiently, requiring consideration of fewer exemplars in determining which acoustic forms were appropriate for the current context, (2) more often involve selection of exemplars that were originally produced in the same syllable currently being produced, and (3) diverge more quickly over time from the original, default forms.

4.1. *Acoustic, linguistic data*

As in Experiment 1, the acoustic memory sequence consisted of a slowly varying multidimensional signal that represented the amplitude envelope of speech across different frequency ranges. It was not feasible to use actual segments or envelope patterns derived from actual productions in this experiment, however. As discussed in Section 2.2, the effects predicted above generally assume that the expected similarity of different-type segments is zero. While this might be true on average within an inventory, the distribution of cross-category similarities is highly nonuniform and asymmetrical, since there are very similar and very dissimilar pairs of sounds. Over the wide variety of syllables and syllable sequences in a language, effects of these nonlinearities on the selection of specific segments in specific contexts would be expected to wash out on average, revealing overall effects of frequency like those seen in Experiment 1. Furthermore, close examination of selection patterns across groups of similar contexts in an appropriately large-scale simulation might be enlightening with respect to the apparent phonotactic and featural relations in a language. Since this experiment was designed to model very slow changes in one set of sounds over time, though, computational (and data) limitations necessitated simulation of a small subset of the entire language. For this model, realistic segment-to-segment similarity asymmetry resulted in artifacts that were large compared to the patterns of interest. Therefore, the acoustic properties of (default versions of) segments in the model were generated using a random process that was closely linked to patterns observed in the corpus but allowed for a more uniform distribution of acoustic content. Likewise, to avoid idiosyncrasies related to correlations between syllable type, content, and position in a word or phrase, syllables were composed of random segments and selected for production based on an artificially created probabilistic grammar that approximated a scaled-down version of a human language.

Twelve unique segments were defined at the beginning of the simulation. Each segment was a 200 ms, four-dimensional sequence representing temporal envelopes in four logarithmically spaced frequency bands ranging from 80 to 8000 Hz. In each dimension, the default envelope for a segment was generated by summing 10 sinusoids with random phase and frequency between 5 and 30 Hz. Individual frequencies were taken from a probability distribution that was created by scaling the modulation spectrum of the corpus (estimated from 100 randomly selected utterances) within this frequency range. Sequences were sampled at 300 Hz and normalized to a total power of 1.0.

These segments were combined at random to form 200 unique three-segment syllables. During the simulation, production was "planned" at the syllable level, based on a probabilistic syllable grammar that dictated that about 60% of the productions would involve the four most frequent syllables. This was intended to represent the fact that the top few percent of syllable types in a language typically represent a majority of total tokens (Levelt & Wheeldon, 1994). It was approximated by first randomly generating a 200×200 syllable transition probability matrix and then simply multiplying four of the rows by 73.5 before normalizing the columns to sum to one. In generating the segmental content of the syllables, a constraint was imposed that each of the segments occurred exactly once in one of the four frequent syllables. This is critical in ensuring that any averaged differences in segment productions relating to syllable frequency are really driven by context and not simply a result of segment frequency, due to segments themselves occurring more often (or only) in one syllable type.

As described in Section 2.1, the segment context comparisons considered both left (acoustic) and right ("linguistic") context information. Linguistic information (the names of segments to be produced) also took the form of a multidimensional signal sampled at 300 Hz with segments of 200 ms length. Twelve dimensions corresponded to the 12 possible segments; a segment name was encoded by adding a rectangular window the length of the segment and total power 1.0 in the appropriate location and dimension and setting the signal to zero elsewhere. Of course, the details of this implementation had no practical effect on the working of the model in the simulations described here, since each segment comparison resulted in either a perfect or zero match. However, it was intended to represent the parallel encoding of abstract (linguistic) and acoustic information in sequences of a similar type that may be compared in a similar way. This similarity will be important in further, more realistic simulations where segment identity and temporal location are represented stochastically.

4.2. *Procedure*

A sequence of 6000 syllables was specified for production, based on the transition probabilities defined by the grammar (the first syllable was selected at random). The model was then prompted to produce the resulting 18 000 constituent segments in order. Production by the model progressed at the segment level (and used only segment-level information) based on selection from previous productions, in the following manner. Each time a segment was produced, a cloud of acoustic exemplars was first generated. Starting with the most recent, each previous production of the target segment was assigned a context-match score, reflecting how similar the context in which the exemplar was originally produced (previous acoustic and following linguistic) was to the current production context:

$$\mathrm{c-match}(t_0, t_e, n_a, n_l, n_e) = \exp\left\{\sum_{d=1}^{D_A} A_{d,t_e-n_a:t_e-1} \cdot A_{d,t_0-n_a:t_0-1} + \sum_{d=1}^{D_L} L_{d,t_e+n_e:t_e+n_e+n_l} \cdot L_{d,t_0+n_e:t_0+n_e+n_l}\right\}$$

where $L_{d,m:n} = (L_{d,m}, \ldots, L_{d,n})^T$ (and analogously for A); A refers to the D_A–dimensional acoustic memory sequence; L is the D_L–dimensional linguistic sequence; t_e and t_0 are beginning indices of the exemplar under consideration and the segment currently being produced; n_a, n_l, and n_e are the lengths of the left (acoustic) and right (linguistic) sequences considered and the length of the target segment. As described above, D_A (the number of frequency bands) was 4 and D_L (the number of segments) was 12; n_a and n_l were both set to two segments, or 120 samples (400 ms) of information.

The matching process proceeded in this way until the summed scores of the tokens considered so far exceeded a constant threshold value. This threshold was set somewhat arbitrarily to 403 ($=\exp(1.5*(n_a+n_l))$), a value that was observed to result in substantial numbers of exemplars (more than a dozen) even in cases where there were very recent close context matches. If the beginning of the memory was reached before this threshold was met, the original, default version of the segment was selected. This occurred only near the beginning of the simulation. Otherwise, an exemplar was selected at random from the cloud, with probability proportional to its context-match score.

Production involved a systematic degradation of the acoustic content of the selected exemplar. As with the acoustic material itself, this degradation process was somewhat arbitrary and abstract, since using, for example, actual coarticulation patterns observed in the database would introduce artifacts related to the specific patterns introduced. It was intended to represent the types and degree of distortions that occur in lenition during speech production: amplitude information is shifted in time, with an eventual potential loss of the highest frequency modulations (due to the sampling rate). Most critically, it had the effect that repeated distortions resulted in progressively lower similarity (as measured by the context match function) of a sequence with its original, default version. That is, the process was unidirectional. This provided a measure of absolute variability or distortion over time, making it possible to compare patterns across time for the same segment in different contexts. The distortion consisted of warping the acoustic pattern for each segment in each dimension slightly to the right (the choice of direction was completely arbitrary), by adjusting the time scale t:

$$\mathrm{warp}(t) = t \times \left[(1-w) + w\frac{t}{n_e}\right]$$

where w is selected—separately for each dimension—from the positive half of a normal distribution with mean 0 and standard deviation 0.01, and then resampling the acoustic pattern according to this adjusted scale.[4] Fig. 5 demonstrates the cumulative effect of multiple applications of the distortion process on one dimension of one segment used in the simulation.

Following this distortion process, the produced sequence was simply appended to the end of A, and the linguistic material was appended to the end of L. Finally, both sequences were multiplied by the constant value 0.999, simulating the decay of memory over time. The result of this decay was that, all other things equal, more recent exemplars were preferred over more distant ones.

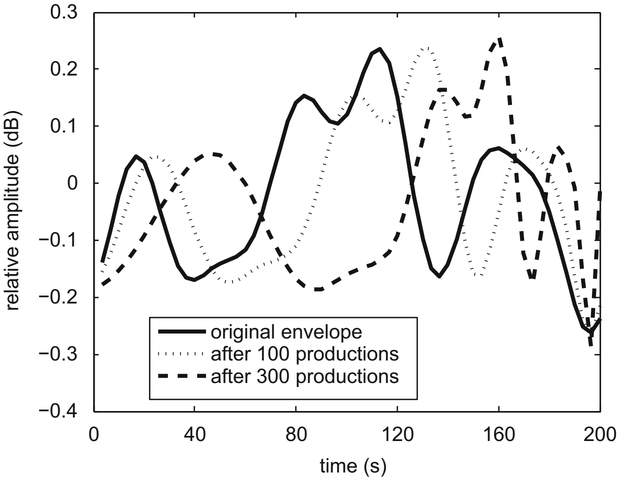

Fig. 5. Sample segment envelope (see text) in default form, after applications of time-warping distortion.

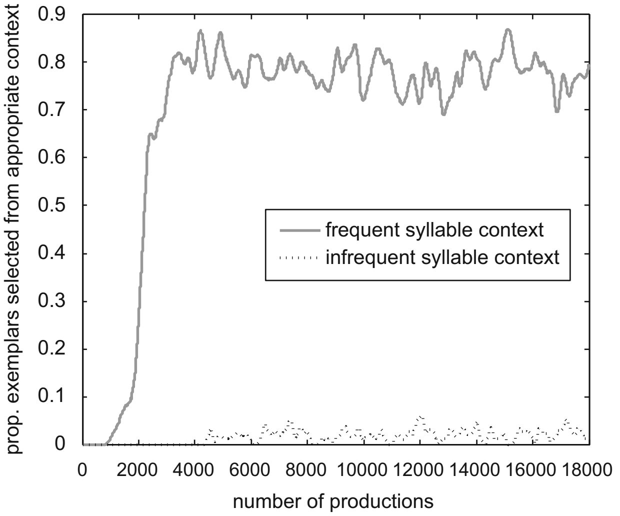

Fig. 6. Proportion of exemplars that were selected from memory sequences involving the same syllable currently under production.

4.3. Results and discussion

Figs. 6–10 summarize the selections made by the model over time. In these figures, the plotted lines represent raw values for each production, smoothed by convolution with a hamming window of length 160 (infrequent) or 240 (frequent) consecutive productions, while the x-axis represents the actual position of the production in the simulation.[5] In all of the figures, there is a transition period of about 5000 productions followed by a more stable period where differences between frequent and infrequent productions can be observed. The initial period corresponds to a stage where the model had insufficient data to make productions based on exemplar selection and relied at least part of the time on the initial, default versions of segments. It might be thought of as representing, in a limited sense, the period immediately after a

[4] A cubic spline interpolation method was used to calculate the values for the 60 samples of the resulting segment representation in each dimension.

[5] This is not identical to smoothing over time, since frequent and infrequent items were interleaved randomly; however, it demonstrates the critical trends in a more detailed manner than, for example, binning data into discrete time blocks. The difference in the length of the smoothing window was required to equate for the typical timing of productions, since frequent syllables accounted for 50% more of the total productions than infrequent syllables.

sound change (or a sound) is introduced in a language, but the effects predicted based on exemplar selection relate mostly to the second, longer period. Critically, none of the productions ever explicitly referenced the syllable or higher levels of organization related to the context, and default segments were never used after about the first 5000 productions.

While syllables were not explicitly referenced, though, exemplars were selected as if they were, at least for the four frequent syllables. Fig. 6 shows the proportion of times that the exemplar that was selected for production originally occurred in the same syllable context as the current production context. In line with prediction (2) above, segments in frequent syllable productions were almost always derived from productions in the same syllable context—effectively resulting in syllable-level productions. These same segments, though, were almost never chosen appropriately for context when the context was less frequent.

Figs. 7–9 demonstrate three important patterns that are closely related to this observation. Fig. 7 shows the size of the exemplar clouds from which segments were selected, Fig. 8 shows the recency (number of productions back) of the exemplar that was actually selected, and Fig. 9 shows the context-match value corresponding to the selected exemplar. Frequent syllable contexts tended to involve smaller clouds (due to higher average context-match scores) that were dominated by a few relatively recent productions of the same syllable. This can be thought of as representing quicker, more efficient selection of exemplars in frequent contexts, since fewer comparisons were required.

As a result, there was a shorter average time between the production of a segment and its “recycling” by selection in another production in frequent compared to infrequent contexts, allowing the acoustic distortion process to compound faster for segments in frequent syllables.

The end result of this compounding process is demonstrated in Fig. 10, which shows the similarity (summed dot product) of segment productions with their original, default forms over time, along with the local variability associated with this match. As shown in Fig. 10, syllables diverge from their default versions faster, and with somewhat more variability, in frequent than in infrequent contexts. It is important to remember that the two plots represent productions of the same six segments in the same average proportions, differing only in the contexts in which they occurred. Thus, it is possible to approximate the effects of syllable-level selection simply by taking local acoustic context into account in a realistic way during segment-level selection, a process that must occur anyway for proper exemplar selection.

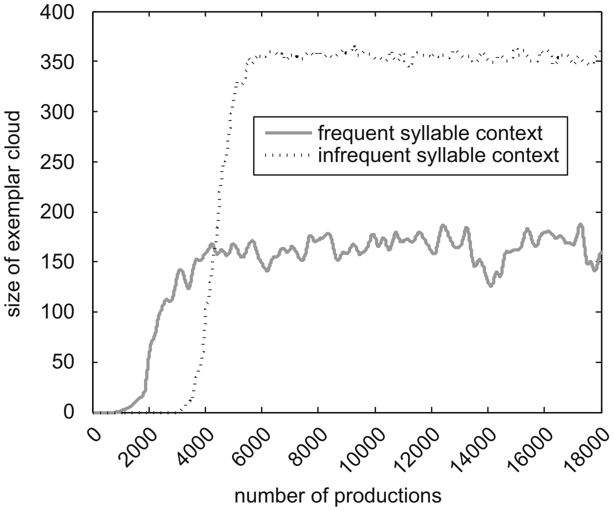

Fig. 7. Size of the exemplar cloud (number of segments) considered for selection.

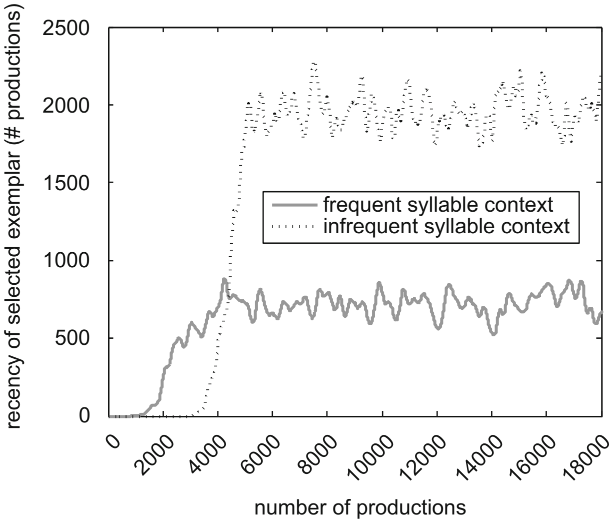

Fig. 8. Age of selected exemplars at the time of production.

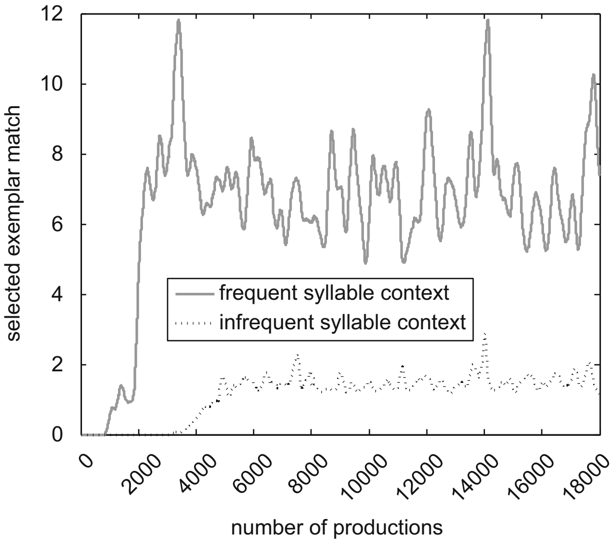

Fig. 9. Absolute context-match value corresponding to selected exemplars.

5. General discussion

It is well known that context is important in characterizing the perception and production of speech sounds. In this study, we set out to quantify this importance in the context of an exemplar-based production process, estimating how much acoustic context must be considered in order to select segment-level tokens that are optimally appropriate for their surroundings (Experiment 1). Making estimates from a large single-speaker corpus, we found that up to about 1 s of surrounding context (0.5 s preceding and 0.5 s following) was useful in determining the acoustic shapes of phoneme categories. We then described a model of how knowledge of typical temporal patterns in speech might guide production, in which segment exemplars are selected one at a time based on their appropriateness for the evolving context, and observed (Experiment 2) that this incremental context-based selection process accounts for observations that are often assumed to implicate an explicit link between discrete segment

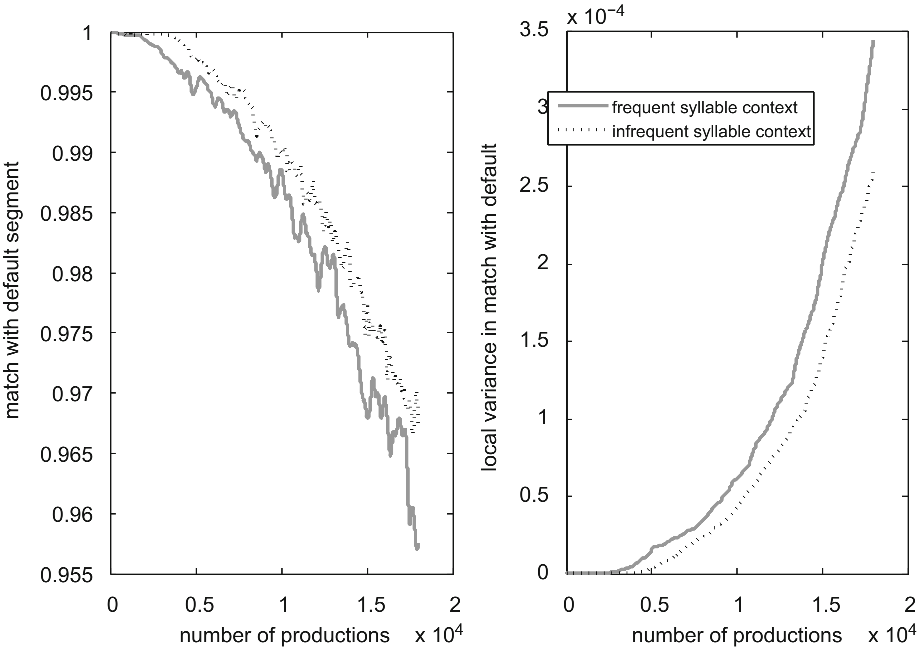

Fig. 10. Correlation of produced segments with the original default versions over time (left) and variance of this measure (right) over a window of 80 (infrequent) or 120 (frequent) productions. Variance was computed normalizing by the sample size (80 or 120), n, rather than $n-1$, so that the overall value was not affected by the different window sizes.

and syllable levels of a linguistic hierarchy. Although there was no reference to a syllable level in the model's selection process, segments produced as part of more frequent syllables were selected more efficiently and gradually took on context-specific patterns, becoming more variable and more affected by lenition processes than the same segments produced in less frequent contexts. This demonstrates that observed syllable frequency effects in production do not necessarily imply "the existence of syllabic units" (Cholin et al., 2006). As such, it provides some support for the notion that usage-based accounts of speech offer more parsimonious accounts of production than models which require discrete levels such as the syllable, since the additional assumptions required by the model (like context-dependent selection) are based on empirical observation. Another important feature of the Context Sequence account is that it provides some (rather strong and probably overly simplistic) hypotheses about the time course of phonetic processing that can be tested experimentally. These two outcomes are discussed further in the following sections.

5.1. Units, levels and categorization in a memory sequence model

The results of Experiment 2 highlight the power of usage-based models in general, demonstrating that segment-level patterns of co-occurrence can account for effects usually assumed to require a syllable level. This is not to say that all linguistic phenomena derive from an acoustic level of analysis; certainly any efficient processing of the speech signal would require reference to categories that span more time than a segment, for example. At the same time, by assuming phoneme category labels and the segment level of representation we have already imported a significant amount of structure from traditional phonological theory. Although this was intended as a computational (and explanatory) convenience, it is important to consider what types of units or levels a memory sequence account like the one described here would actually require. We do not have sufficient data to propose a detailed theory, but will suggest that the same memory sequence might be referenced simultaneously by category labels at different time scales, and that segmentation in general might be more data-driven than the result of a predefined hierarchy. Some pilot experiments suggest that a useful segmentation approach may follow straightforwardly from the types of direct signal envelope comparisons described in Experiment 1. We have considered match sequences similar to those portrayed in Figs. 2 and 3, but where the total match of a probe sequence with an entire memory is measured as a function of the length of the sequence as more and more information is added to its right side. The derivative of such a function can be thought of as a measure of how likely or predictable each bit of new information is given the preceding portion of the probe sequence, and its inverse might be considered a measure of "boundariness," or how likely each point is to represent a juncture between functional units with separate category labels. Since transitions between acoustic events that occur at traditionally described (word, syllable, phrase, etc.) boundaries are probably less likely than those occurring within these units (cf. Hay et al., 2003; Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1998; Saffran, Newport, Aslin, Tunick, & Barrueco, 1997), we thus considered what type of segmentation might result from simple acoustic comparisons in context. In a variety of experiments using both envelope representations like those discussed here and more symbolic sequences involving phonological features, and considering data from several different speech corpora, we generally observe that at least syllable and word boundaries tend to coincide with local maxima in "boundariness" sequences defined in this way. Further research will be required to determine whether this type of segmentation, along with a complementary categorization process, could result in a comprehensive Context Sequence-like account of phonological, and perhaps higher-level, processes. These results do suggest, though, that if "holistic gestural scores" are in fact remembered and used in the production of speech (Levelt & Wheeldon, 1994), such patterns may not need to reference the syllable, or even the whole word or other units that have a strictlyhierarchical origin (e.g. Levelt et al., 1999). Although they might often coincide with these levels, their independent status could be driven by the repeated production of sequences of gestures more generally.

Also related to the issues of segments and segmentation, the Context Sequence model as described here does rely on the presence of categories and category labels—even if they do not (always) correspond to invariant, traditionally assumed levels such as the segment or syllable. This of course implies mechanisms for forming categories and for assigning category labels to newly perceived exemplars. It was not necessary to specify such mechanisms in the simulations described here, since they only involved selection of existing categories from memory. In providing a framework for taking context into account during exemplar comparison, though, the Context Sequence account does have important implications for existing approaches to categorization and learning. A central question in second language phonetic research, for example, is how categories from multiple languages that occupy similar or overlapping regions of acoustic space can be simultaneously maintained (identified and produced successfully) by language users (e.g. Flege, 1995). The assumption that sounds are always considered as part of a surrounding context provides a framework for studying some aspects of this question, and results in some concrete predictions. The success with which a non-native contrast can be made, for example, should depend not only on the similarity of the sounds involved to each other and to native categories and the speaker's overall proficiency, but also on the speaker's exposure to the specific context in which the L2 sound is being presented, and on the distinctiveness of contexts in which the sounds typically occur from sequences in the speaker's L1.

5.2. Model parameter settings, context frequency differences and diffusion

The results of both simulations were generally robust to a range of settings regarding stimulus specifications and encoding, parameter values, etc. In Experiment 2 for example, moderate changes in the numbers of syllables and segments, the match criteria, and the forgetting and warping constants in the model affected the time trajectory of the trends seen in Figs. 6–10 in predictable directions (e.g. larger inventories or less warping caused the patterns to be stretched in time), but the overall pattern of effects remained the same. One trend that was surprisingly stable across different settings was the small size of the frequency-based acoustic divergence and variability effects (as in Fig. 10) relative to the other frequency effects observed in Experiment 2. This was because the effect shown in Fig. 10 was an indirect product of the selection of more recent, more context-appropriate exemplars in frequent syllables and not a deterministic process. There was substantial "cross-pollination" of segmental acoustic forms across syllable contexts, mostly from frequent to infrequent contexts, causing the divergence from the default form in infrequent contexts (left panel of Fig. 10) to progress much more quickly than it would have if all of the exemplars had been selected from infrequent original contexts. It was possible to encourage selection from infrequent contexts for these productions by greatly increasing the influence of very close context matches (for example by scaling the context-match score exponentially), which in turn dramatically delayed the degradation process for infrequent context since selected exemplars tended to be very old ones. A model modified in this way seems unappealing, though, for both theoretical and empirical reasons. First, it makes the model's reliance on acoustic rather than syllable context—and the resulting differences in lenition over time—rather trivial, since it essentially enforces selection at the syllable level rather than allowing it to emerge from the statistics of a more variable selection process. Second, observed frequency effects in production, except for some extremely frequent sequences (e.g. ə̞̃ ə̃ ə̝̃ for "I do not know," Hawkins, 2003) that the current data were not intended to represent, *are* typically relatively small compared to the total range of variation with which segments may occur. What was observed in the present model might be thought of(in a limited sense) as something like lexical diffusion, in that alternations introduced and primarily driven by a subset of (frequent) syllables gradually disseminate across the syllable inventory. The important difference here is that it is only local context consideration, and not syllable or word identity, that drives the acceleration and more generally dictates where changes are more likely or appropriate. In any case, larger simulations with more frequent and infrequent syllable (and word, etc.) types will have to be considered to determine whether this could account for diffusion and related processes actually observed in language change.

5.3. Other limitations of the context sequence account

Apart from not specifying identification and category-formation processes, the Context Sequence model as described here is an incomplete account of phonetic competence in several important respects. First, the simulation in Experiment 2 only addressed the dynamics of how knowledge of the acoustics of speech sounds influences production, without referencing the actual articulation process. This was not intended to suggest that phonetic knowledge is primarily acoustic or auditory in nature (cf. e.g. Diehl & Kluender, 1989; Levelt et al., 1999); it will be necessary to address how articulatory and motor processes relate to or are integrated with a selection process like the one we have outlined (Guenther & Perkell, 2004). A coherent account of these relationships will be necessary in order to evaluate the Context Sequence selection process more completely, and will likely require the use of articulatory as well as acoustic corpora.

Similarly, following previous exemplar-based production models (e.g. Pierrehumbert, 2001), we have treated acoustic variability related to lenition and other sources simply as a unidirectional probabilistic distortion of the segmental acoustic form, ignoring the multiple, interacting sources of this variability in actual speech production (articulation constraints, social influences, desire to maintain contrasts, etc., e.g. Lindblom, 1990). One obvious shortcoming of this approach is that the repeated application of a distortion process such as the one depicted in Fig. 5 would eventually lead to more and more variable categories over time. Pierrehumbert (2001) addresses this problem by implementing a category entrenchment mechanism whereby productions are based on weighted combinations of several existing exemplars rather than a single exemplar. The addition of such a feature might make the results of Experiment 2 more realistic (i.e. the increased variance over time shown in Fig. 10 would be decelerated while the frequency differences would persist). On the other hand, a more realistic representation of the constraints on the distortion process itself (for example, by introducing an additional process that deflects the exemplar from neighboring categories) might also contribute to the preservation of sharp category distributions during language change.

Relatedly, we have modeled context effects in production simply as probabilistic selection of exemplars based on context similarity, ignoring the separate contributions of mechanical constraints on articulation and more arbitrary, language-specific patterns of variability in determining "context appropriateness". While we would be careful not to implicate any model of phonetic knowledge in describing effects that are readily explained by physical mechanisms (e.g. Browman & Goldstein, 1992), it may not be interesting or even possible to separate these different influences in specifying the acoustic output, since they work and probably evolved in parallel. At the very least, it is readily observable that languages differ in the types and degree of—and perceptual compensation for—coarticulation (Beddor, Harnsberger, & Lindemann, 2002). Thus, while some aspects of context-dependence are certainly deterministic, phonetic knowledge of the type modeled here is necessary to determine which processes lead to actual language change.

Finally, one dimension of the speech signal that is often discussed as "indexical" information very likely to be stored in detail is fundamental frequency (Goldinger, 1996). Since the envelope-based acoustic representations considered here do not encode this property (or many others), it could certainly be criticized as not including all of the acoustic details that are needed to specify phonetic exemplars. This probably had little effect on the simulations described here, since they were mostly concerned with productions of a single speaker, but it is almost certain that f0 and other, fine structure information play a role in detailed memory for speech.

References

Beddor, P. S., Harnsberger, J. D., & Lindemann, S. (2002). Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates. *Journal of Phonetics*, *30*, 591–627.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, *23*, 291–320.

Browman, C., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*, 155–180.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically-conditioned sound change. *Language Variation and Change*, *14*, 261–290.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*, 529–551.

Cholin, J., Levelt, W., & Schiller, N. (2006). Effects of syllable frequency in speech production. *Cognition*, *99*, 205–235.

Diehl, R., & Kluender, K. (1989). On the objects of speech perception. *Ecological Psychology*, *1*, 121–144.

Eagleman, D., Tse, P., Janssen, P., Nobre, A., Buonomano, D., & Holcombe, A. (2005). Time and the brain: How subjective time relates to neural time. *The Journal of Neuroscience*, *25*, 10369–10371.

Flege, J. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 229–273). Timonium, MD: York Press.

Foulkes, P., & Docherty, G. (2005). The social life of phonetics and phonology. *Journal of Phonetics*, *34*, 409–438.

Goldinger, S. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology*, *22*, 1166–1183.

Goldinger, S. (1997). Words and voices: Perception and production in an episodic lexicon. In K. Johnson, & K. Mullenix (Eds.), *Talker variability in speech processing* (pp. 33–66). San Diego: Academic Press.

Goldinger, S. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279.

Guenther, F., & Perkell, J. (2004). A neural model of speech production and its application to studies of the role of auditory feedback in speech. In B. Maassen, R. Kent, H. Peters, P. Van Lieshout, & W. Hulstijn (Eds.), *Speech motor control in normal and disordered speech* (pp. 29–49). Oxford: Oxford University Press.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*, 373–405.

Hay, J., Nolan, A., & Drager, K. (2003). From fush to feesh: Exemplar priming in speech perception. *Linguistic Review*, *23*, 351–379.

Holt, L. (2005). Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychological Science*, *16*, 305–312.

Holt, L., & Wade, T. (2004). Non-linguistic sentence-length precursors affect speech perception: Implications for speaker and rate normalization. In *Proceedings of from sound to sense*: *Fifty+ years of discoveries in speech communication*.

Hooper, J. (1976). Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (Ed.), *Current progress in historical linguistics*. Amsterdam: North-Holland.

Jeffress, L. (1948). A place theory of sound localization. *Journal of Comparative and Physiological Psychology*, *41*, 35–39.

Jescheniak, J., & Levelt, W. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 824–833.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & K. Mullenix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego: Academic Press.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*, 485–499.

Kluender, K., Coady, J., & Kiefte, M. (2003). Sensitivity to change in perception of speech. *Speech Communication*, *41*, 59–69.

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), Barcelona.

Ladefoged, P., & Broadbent, D. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*, 98–104.

Levelt, W., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Levelt, W., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary?. *Cognition*, *50*, 239–269.

Licklider, J. (1951). A duplex theory of pitch perception. *Experientia*, *7*, 128–134.

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, *35*, 1773–1781.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the h and h theory. In W. J. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Kluwer.

Lindblom, B., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, *42*, 830–843.

Lively, S., Logan, J., & Pisoni, D. (1993). Training Japanese listeners to identify English /r/ and /l/. ii. The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242–1255.

Loizou, P., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *Journal of the Acoustical Society of America*, *106*, 2097–2103.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, *103*, 18866–18869.

Mann, V. A., & Repp, B. H. (1980). Influence of vocalic context on perception of the /sh/-/s/ distinction. *Perception and Psychophysics*, *28*, 213–228.

Mauk, M., & Buonomano, D. (2004). The neural basis of temporal processing. *Annual Review of Neuroscience*, *27*, 304–340.

Miller, J., & Liberman, A. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception and Psychophysics*, *25*, 457–465.

'Ohman, S. (1965). Coarticulation in vcv utterances: Spectrographic measurements. *Journal of the Acoustical Society of America*, *39*, 151–168.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: Benjamins.

Pisoni, D. (1997). Some thoughts on 'normalization' in speech perception. In K. Johnson, & K. Mullenix (Eds.), *Talker variability in speech processing*. San Diego: Academic Press.

Port, R. (2007). How are words stored in memory? Beyond phones and phonemes. *New Ideas in Psychology*, *25*, 143–170.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Saffran, J., Johnson, E., Aslin, R., & Newport, E. (1998). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*, 27–52.

Saffran, J., Newport, E., Aslin, R., Tunick, R., & Barrueco, S. (1997). Incidental language learning: Listening—and learning—out of the corner of your ear. *Psychological Science*, *8*, 101–105.

Schweitzer, A., Braunschweiler, N., Klankert, T., Möbius, B., & Säuberlich, B. (2003). Restricted unlimited domain synthesis. In *Proceedings of Eurospeech-2003* (pp. 1321–1324), Geneva.

Schweitzer, A., & Möbius, B. (2004). Exemplar-based production of prosody: Evidence from segment and syllable durations. In *Proceedings of the speech prosody 2004 conference* (pp. 459–462), Nara.

Shannon, R., Zeng, F., Kamath, K., Wyngonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Stevens, K. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America*, *111*, 1872–1891.

Summerfield, Q. (1980). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1074–1095.

Wade, T. (2007). Implicit rate and speaker normalization in a context-rich phonetic exemplar model. In *Proceedings of the 16th international congress of phonetic sciences*, Saarbrücken.

Wade, T., & Möbius, B. (2007). Speaking rate effects in a landmark-based phonetic exemplar model. In *Proceedings of interspeech 2007*, Antwerp.

Walsh, M., Schütze, H., Möbius, B., & Schweitzer, A. (2007). An exemplar-theoretic account of syllable frequency effects. In *Proceedings of the 16th international congress of phonetic sciences*, Saarbrücken.

Walsh, M., Schütze, H., Wade, T., & Möbius, B. (2007). Accounting for phonetic and syntactic phenomena in a multi-level competitive interaction model. In *ESSLLI workshop on exemplar based models of language acquisition and use*, Dublin.

Whalen, D. (1990). Coarticulation is largely planned. *Journal of Phonetics*, *18*, 3–36.

Zeng, F., Nie, K., Stickney, G., Kong, Y., Vongphoe, M. Bhargave, A., et al. (2005). Speech recognition with amplitude and frequency modulations. *Proceedings of the National Academy of Science*, *102*, 2293–2298.