Slightly revised version of paper in *Proc. Interspeech 2006*

A Model of the Regularities Underlying Speaker Variation: Evidence from Hybrid Synthesis

Susan R. Hertz

NovaSpeech LLC and Cornell University
Ithaca, NY, USA

ABSTRACT

This paper presents the framework of a speech model, tentatively called the "hybrid model," which offers an explanation of how listeners can identify phonemes in an incoming speech signal despite the vast amount of cross-speaker and contextual variation. Fundamental to the model are two basic speech units into which listeners process the incoming speech stream: acoustic consonant clusters and acoustic nuclei. Acoustic nuclei are responsible for speaker identity, but acoustic consonant clusters are more generic and can even be substituted across speakers with negligible impact on speech quality. The paper focuses on acoustic consonant clusters, showing that much of the variability in them is perceptually irrelevant, and how the hybrid model accounts for listeners' ability to parse them into phonemes. The paper supports the model as applied to English by drawing on experiments in hybrid synthesis, a technique in which speech is produced by splicing together segments from different speakers (natural or synthetic) [1].

Index Terms: speech perception model, speech synthesis, speech variation, consonant perception, speaker identification

1. Introduction

One of the greatest challenges facing speech researchers is the highly variable nature of the speech signal. Even the same speaker might render the same intended utterance in decidedly different ways on separate occasions. When different speakers produce an utterance (even with the same intended prosodic characteristics), there is a virtual certainty that their renditions will have many acoustic dissimilarities. Differences can result from many factors, such as individual vocal tract physiologies and production strategies. In spectrograms of the same utterance produced by different speakers, for example, one might find that one speaker has realized an unstressed syllable with a clearly identifiable reduced vowel, while another has realized the same syllable simply by extending the last consonant of the preceding syllable; or that one speaker has rendered a phonological stop with a clear closure, but there is no hint of such closure for another. To complicate matters further, the same acoustic segment might correspond to one phoneme in one context, but to a different phoneme in another. The [p] of *speech*, for instance, is acoustically similar to the [b] of *beach*, as evidenced by the fact that *speech* sounds like *beach* if the initial [s] noise is eliminated.

Just how humans manage to make sense of the highly variable and complex speech signal has been the subject of much debate. Researchers don't even agree on questions as fundamental as whether listeners parse an incoming speech stream into basic units like phonemes in a rule-governed fashion. An alternative view, for example, held by a growing number of researchers, is that listeners extract phonological units like phonemes and words by comparing the incoming speech stream against thousands of remembered tokens ("exemplars") of such units acquired through their listening experience [2].

Our work suggests that, particularly in the case of consonants, **speech is much more rule-governed than might appear at first glance**. Although consonant phoneme sequences are often realized in acoustically different manners by two speakers, they nevertheless share certain robust characteristics through which listeners recover the intended phonemes. Empirical support for this view is offered by our recent experiments in hybrid synthesis, in which we successfully substituted, in principled ways, consonants spoken by one speaker with those of others (both natural and synthesized) with little or no impact on speaker identity, phoneme intelligibility, or naturalness. More specifically, this work has shown that:

- Over half of the segments (comprising 60-70% of the duration of a typical English utterance) can be replaced by segments from another speaker with little perceptual effect.
- The "surrogate segments" can often have decidedly different acoustic characteristics than the segments they replace, and be taken from speakers of the opposite gender, very different ages, and markedly different vocal characteristics.
- Even formant-synthesized segments produced with general rules can serve as surrogates with little or no perceptual degradation to the resulting speech.

While these results might seem to defy conventional logic, and are certainly at odds with the type of evidence presented by proponents of exemplar-based models, they can all be easily accounted for by the hybrid model. The first section that follows outlines this model, while the second presents evidence for it from our hybrid synthesis experiments.

2. Hybrid model

The hybrid speech model has evolved from many years of research in multi-language speech synthesis by the author and her collaborators using a perceptually- and linguistically-oriented approach [3-5]. It is an outgrowth of Hertz's earlier phone-and-transition (P&T) model of speech [3,4], which served as the basis for the language-universal components of the multi-voice ETI-Eloquence formant-based synthesis rules for thirteen languages [5]. In the P&T model, separate phone and transition units (based primarily on F2 behavior) are posited to account for acoustic and perceptual phenomena, such as the stability of certain formant transition durations relative to phones, and the fact that stop aspiration tends to align precisely with transitions [3].

In the P&T model, phones and transitions are grouped into higher level units to account for various acoustic and perceptual patterns [4]. Of particular importance is the acoustic nucleus (AN), which is the part of the waveform that corresponds to the syllable nucleus. The AN for English consists of the vowel phone of the syllable, any following tautosyllabic sonorants, and any voiced portions of transitions on the outside edges of the sequence (henceforth "edge transitions"). (The status of high-

amplitude tautosyllabic sonorants before the vowel is under investigation.) With this unit a variety of timing patterns can be captured, including the trading relation among the edge transitions and phones in the nucleus, with phones stretching and shrinking to accommodate the contextually more stable edge transition durations [3].

The hybrid model posits that not only ANs, but also acoustic consonant clusters (ACCs) play a central role in speech organization. ACCs are the portions of speech between ANs, consisting of any sequence of consonant phones, any intervening transitions, and any devoiced portions of transitions at the edges of the sequence. Heavily reduced vowels, which coarticulate strongly with adjacent segments, have no inherent F2 targets of their own, and whose amplitudes are lower than full vowels, are assumed to pattern with ACCs in the ensuing discussion.

During many informal gating and cross-splicing experiments we have conducted in several languages, we have continually noticed that shortly after the initial edge transition of an AN has been processed by the listener, the phonemes of the preceding ACC can be unambiguously determined. Note that the contextually robust nature of the initial edge transition is a key ingredient enabling a stable parsing point early in the AN.

Consider, for example, the phrase *speech parsing*. The waveform corresponding to /sp/ would be queued up by the listener until the acoustic nucleus [i] is encountered (at the sudden abrupt rise in energy and the relatively periodic waveform). A short distance into this nucleus (on the order of 30 ms), after the *durationally stable* labial transition has been processed, the listener would parse the ACC into /sp/, using its *gross relational* acoustic patterns as well as general characteristics of the edge transition. For example, the listener would use the facts that [s] in this context (for all speakers) has a relatively long period of high intensity noise at relatively high frequencies, [p] has a period of silence followed by a low energy diffuse burst, and the F2 in the edge transition rises. Note that the shape of this edge transition does not uniquely cue the labial place of the preceding stop, since an alveolar transition before a front vowel would have a similar shape. However, listeners don't parse the edge transition independently of the preceding ACC; thus the markedly different spectral and amplitude structures of labial and alveolar stop bursts help differentiate labial and alveolar consonants before front vowels.

Whenever ACCs are processed, any intervening syllable boundaries, like the one between the phones [čp] of *speech parsing*, are also determined using general, speaker-independent acoustic patterns and phonological constraints. For example, the different syllable structure of *loose peaks* vs. *Lou speaks* is identified in part by the relatively greater amount of aspiration of the [p] of *peaks*. In *loose spoke* vs. *Lou spoke*, the longer [s] noise duration in the first case signals that there are two /s/ phonemes.

In general, timing relations within and across ACCs are critical for their decomposition into phonemes. For instance, when lenitions occur within an ACC (e.g., /nts/ → [ns]) neighboring segments often compensate to provide stable durations across combinations of segments within the cluster (in the [ns] variants, the [n] is longer than in the [nts] variant). Similarly, the main perceptually salient cues to the [s] of [asa] are not only its spectral shape and neighboring edge transitions, but also its relatively long duration. If the duration is shortened beyond some threshold, the [s] sounds like /z/. When shortened further to the duration typical of [d] in this context, it sounds like /d/. Moreover, acoustically similar segments that have similar edge transitions, such as a low intensity labial fricative [f] and a non-aspirated labial stop [p], have systematically and markedly different durations. In short, it appears that timing patterns in languages are strategically organized to enhance phonological contrasts, and both speakers and listeners adhere to these organizational principles.

Note that it is not surprising to find that timing patterns play such a central role in phoneme identification, since all speakers can produce them in similar ways regardless of vocal tract characteristics, and durational cues are also more robust in noisy environments than are many spectral cues, a point also made in [6]. Despite their central role, however, timing relations are often overlooked in quests to find robust cues to phoneme realizations. While durational cues tend to be neglected, unwarranted weight is often given to perceptually irrelevant spectral details. It is important to keep in mind that **an observed event is not necessarily a perceived event.** Our hybrid experiments strongly suggest that listeners abstract away from acoustic differences resulting from individual vocal physiologies when parsing ACCs into phonemes. It is reasonable to posit further that variation in production that is under the speaker's control is permitted only when such variation will not disrupt the perceptual relations required to discern the phonemes.

3. Hybrid experiments

The hybrid model has grown out of a general experimental paradigm of iterative hypothesis formulation and testing. We run both formal and informal perceptual tests; evaluate the results; try to correlate the scores or perceived problems with the phonological and phonetic characteristics of the stimuli; revise our hypotheses accordingly; and generate new stimuli to test the revised hypotheses. These cycles of exploration and experimenttation are so frequent that rarely do two listeners hear exactly the same stimuli. However, we are careful in formal experiments to obtain enough judgments on a core set of stimuli so as to be able to validate our hypotheses through statistical measures. Unless mentioned otherwise, all results presented in this paper for hybrid synthesis were statistically significant ($p < .05$).

In our most recent experiments, we elicited speaker identification (SI) and speech quality (SQ) judgments for a variety of hybrid stimuli for the sentences in Table 1. These sentences contain several sequences where we would expect considerable variation in production across speakers, including function words, unstressed syllables, and syllable-final consonants.

The expert skier agreed to tee up with the pro golfer.
Sheila forgot that Thursday's the day before Friday.
Monica Naimoo never knew Bonnie's mother.
Computers with multiple voices are incredibly cool.

Table 1: Experimental utterances

For each of these sentences we constructed stimuli in which we mixed speech from 11 different speakers collectively called the *hybrid speakers*. The hybrid speakers included eight human speakers and three synthetic speakers. The eight human speakers included three children ages six to eleven, two young adults in their twenties, and three adults in their fifties.

Six of the human hybrid speakers, described in Table 2, functioned as *target speakers* whom we aimed to mimic with the hybrid stimuli. Each of the target speakers recorded an 82-word passage, arbitrarily selected from a novel, which was used to train listeners on their voices for the SI experiments.

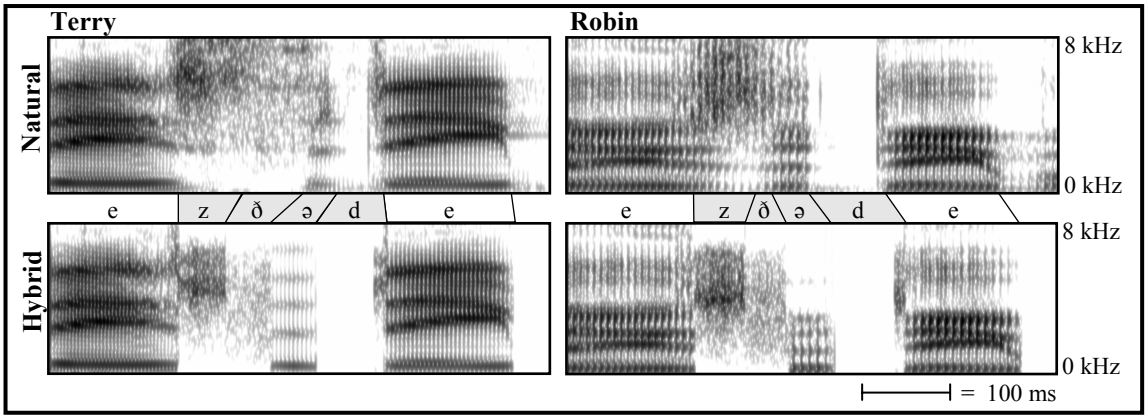

Figure 1: Natural and hybrid versions of *Sheila forgot that Thursday's the day before Friday* for two speakers

The synthetic speakers were produced with a state-of-the-art rule-based formant synthesis system based on ETI-Eloquence [5]. This system has shown that, contrary to conventional wisdom, highly intelligible (and, as shown below, natural and mimetic) consonants can be produced with simple, perceptually-oriented rules that set just a few contextually appropriate values for just a few parameters. (A "mimetic" segment is one that can be used to mimic a particular speaker.)

In addition to rule-based formant synthesis, for the production of some synthetic surrogates, we also used a technique we call "model-based formant synthesis" in which we synthesized segments in accordance with our hypothesized rules or principles. Model-based synthesis should not be confused with copy-based formant synthesis, in which the specific details of particular utterances are copied, a technique we also used for selected stimuli. For production of the waveforms with all three types of formant synthesis, a KLSYN-88 style synthesizer was used [7].

Name	Sex	Age
Morgan	M	11
Terry	F	11
Pat	M	23
Nat	F	24
Robin	M	54
Lee	F	54

Table 2: Target speakers

For purposes of constructing the hybrid stimuli for the SI and SQ tests, we labeled the synthetic and natural utterances into ANs and ACCs, and each ACC into perceptually-important smaller units. To create the core stimuli for both tests, we made a variety of different types of substitutions into copies of 18 base utterances consisting of the four test sentences spoken by the six target speakers. In most of these stimuli, all consonants were replaced by those of another speaker, while in a few, only selected segments, such as reduced vowels or sibilants, were replaced.

In accordance with earlier hybrid results [8], F0 values in voiced surrogates were interpolated between the F0 target in preceding and following voiced segments; non-sonorant clusters were generally substituted across speakers as whole chunks; voiced sonorant surrogates were produced via model-based formant synthesis; amplitudes of surrogates were adjusted when necessary based on general principles to stand in appropriate relations to neighboring segments; and durations were taken from the surrogates, unless these sounded unnatural, in which case they were adjusted. Since the vast majority of the hybrid stimulus surrogates were taken from rule-based formant synthesis, durations sometimes had to be adjusted simply due to imperfections in the synthesis rules. Interestingly, however, when a synthetic surrogate duration was adjusted for one speaker, that same duration could then generally be used for all the stimuli, regardless of target speaker.

Figure 1 shows representative spectrograms for the highlighted fragments of two natural and two hybrid renditions of *Sheila forgot that Thursday's the day before Friday*. The spectrograms in the top row show natural versions for the female child Terry and the older adult male Robin. One can clearly see different formant and noise frequencies resulting from the speakers' different vocal tract sizes as well as different voicing patterns in the fricatives resulting from their different production strategies.

In the hybrid sentences in the bottom row, formant-synthesized surrogates of equal duration were substituted for both speakers. These were produced with rule-based formant synthesis (adult female voice for Terry, adult male voice for Robin) for all segments except [ə], which was produced with model-based formant synthesis to produce formant values plausible for the context and general characteristics of the target speakers being modeled. Despite the clear differences between the original and surrogate segments, and the similar surrogates used for both cases, the hybrid sentences were considered highly mimetic and natural, as discussed below.

3.1 Speaker identification results

In the SI experiments, 34 listeners each characterized 45 stimuli selected from a cohort of 123 total stimuli in terms of whether they sounded like one of the target speakers and how much so (where 1 = "exactly like," 2 = "a lot like," 3 = "similar to," and 4 = "a bit like"), or, for those stimuli that didn't sound like a target speaker, in terms of gender and one of four predefined age groups. Nineteen of the listeners were familiar with at least one of the hybrid speakers.

In addition to 55 core hybrid stimuli, the experiment included 23 natural utterances, six produced via copy-based formant synthesis, nine produced by rule-based formant synthesis, and 13 produced by two state-of-the-art synthesis systems that generate speech by means of corpus-based waveform concatenation (CBWC), in which waveform fragments from a single speaker are selected from a corpus and pieced together to produce an utterance. CBWC1 is a system reputed for its natural-sounding voice quality, while CBWC2 is better known for its low memory usage. Collectively, the stimuli represented 22 different voices, including non-target voices of ages and genders similar to those of the target voices. Listeners could play each stimulus as often as desired. Before beginning the experiment, listeners were trained on the target speakers' identification paragraphs, and

asked to characterize each speaker in terms of gender and age.

The results were striking: Listeners familiar with the hybrid speakers correctly identified them 96% of the time, and considered them to sound a lot like the target speakers, giving them an average similarity score of 1.99. These results are in line with those for the natural speech tokens, which were correctly identified 98% of the time with an average score of 1.37. What's more, it was clear from their comments that listeners had no idea that the voices they heard consisted of more than one speaker!

Unfamiliar listeners correctly identified both hybrid and natural stimuli less accurately, but again the results for both classes of stimuli were similar (hybrid: 78%/2.17; natural: 79%/1.97). Age and gender identification results were comparable as well.

Rule-based formant synthesis was judged significantly less accurately by all listeners than hybrid stimuli, with only 75% correct gender identification compared with 97% for hybrid stimuli. Consonant surrogates taken from the misidentified rule-based synthesis, however, did not degrade speaker identification, a fact lending strong support to their cross-speaker generality.

3.2 Speech quality results

In the SQ experiments, listeners were asked to judge the overall naturalness of the stimuli and mark specific problems. Thirty-four listeners participated, characterizing approximately 60 stimuli (of types similar to those in the SI experiments) selected from a cohort of 143 total stimuli in terms of their overall naturalness and specific problems. Listeners were first trained on a representative range of stimuli of varying qualities, as determined in a pilot experiment, so they would have a basis for their judgments.

After playing a stimulus once, listeners were asked to rate the overall naturalness on a five-point scale, where 1 = "very natural," 2 = "fairly natural," 3 = "mid-range," 4 = "fairly unnatural," and 5 = "very unnatural." Next they played the stimulus as often as desired to mark problems on individual words or on the whole utterance, using categories such as *non-human-sounding*, *unexpected pronunciation*, *bad rhythm*, *foreign accent*, *nasal-sounding*, *hard to understand, speech impediment,* and *choppy-sounding*.

Table 3 shows the overall naturalness results for each type of stimulus as well as the number of tokens and responses for each type (CBFS means "copy-based formant synthesis" and RBFS means "rule-based formant synthesis"). A one-way ANOVA revealed significant differences between groups ($F = 439.75$, $p < .001$). Posthoc comparisons using the Bonferroni correction showed all means to be significantly different from each other except for CBWC2 and CBFS ($p < .001$). The average score for all of the core hybrid stimuli was 1.97. The average scores for the specific hybrid utterances in Figure 1 were 1.70 and 1.85 respectively, based on 20 listeners.

Type	Score	#Tok	#Rsp
Natural	1.40	23	295
Hybrid	1.97	92	1265
CBWC1	2.85	11	266
CBWC2	4.04	2	26
CBFS	4.21	6	72
RBFS	4.84	9	187

Table 3: Naturalness scores

Despite the relatively few tokens, a comparison of CBFS vs. CBWC1 is interesting. The F0, timing, and spectral patterns for the copy-based tokens were copied directly from natural utterances, yet the stimuli were judged as quite unnatural (4.21). The CBWC1 stimuli, in which all units were taken from the same speaker but were not necessarily contextually appropriate, were judged as considerably more natural (2.85) than the copy-based stimuli. Further, while copy-based stimuli were generally marked as non-human-sounding, CBWC1 stimuli were often marked as containing words with foreign accents, unexpected pronunciations, and unnatural timing. **These differences strongly suggest that one of the primary correlates of the overall quality scores is whether the AN comes from a human speaker or not.** Once again, we see that ANs and ACCs are very different beasts.

4. Conclusion

This paper has presented the results of recent hybrid experiments involving speaker identification and speech quality judgments, which demonstrate the cross-speaker generality of ACCs, and, implicitly, the role of ANs in cuing individual voice quality. The results fit neatly within the hybrid model of speech organization outlined at the onset, which posits that humans organize both speech production and perception around these units. The paper has presented evidence that listeners abstract away from cross-speaker variation when processing ACCs, and that cross-speaker generalizations, including timing relations, account for listeners' ability to parse these units into phonemes. While the success of hybrid synthesis has been surprising to some, we view it as a natural consequence of the hybrid model, and as lending strong support to its basic premises. The paper has focused on application of the model to English; however, preliminary results suggest that it extends to other languages as well.

5. Acknowledgements

The author is grateful to Isaac Spencer for tireless efforts on the hybrid experiments. Thanks also to Isaac, Draga Zec and Jessica Sharkness for helpful suggestions on drafts of this paper. This work was supported in part by NIH Grant R43 DC006761-01. The views are those of the author and do not necessarily reflect those of the agency.

6. References

[1] Hertz, S.R., 2002. Integration of rule-based formant synthesis and waveform concatenation: a hybrid approach to text-to-speech synthesis. *Proc. IEEE Workshop on Speech Synthesis.*

[2] Pierrehumbert, J., 2000. Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee and P. Hopper (eds.) *Frequency effects and the emergence of linguistic structure*, Amsterdam: John Benjamins, 137-157.

[3] Hertz, S.R., 1991. Streams, phones, and transitions: toward a phonological and phonetic model of formant timing. *J. of Phonetics* 19, 91-109 *Special Issue on Speech Synthesis and Phonetics*, ed. by R. Carlson.

[4] Hertz, S.R. and Huffman, M.K., 1992. A nucleus-based timing model applied to multi-dialect speech synthesis by rule. *Proc. ICSLP* 2, 1171-1174.

[5] Hertz, S.R., Younes, R.J., Zinovieva, N., 1999. Language-universal and language-specific components in the multi-language ETI-Eloquence text-to-speech system. *Proc. 14th ICPhS,* 2283-2286.

[6] van Wieringen, A. and Pols, L., 2006. Perception of highly dynamic properties in speech. In S. Greenberg and W. A. Ainsworth (eds.) *Listening to Speech: An Auditory Perspective*, New Jersey: Lawrence Erlbaum Assoc., 21-38.

[7] Klatt, D.H. and Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *JASA* 87**,** 820-857**.**

[8] Hertz, S.R. and Goldhor, R., 2004. When can speech segments serve as surrogates? *Proc. From Sound to Sense: 50+ Years of Discoveries in Speech Communication* (poster available at http://www.novaspeech.com).