WHEN CAN SEGMENTS SERVE AS SURROGATES?[1]

Susan R. Hertz[§], Isaac C. Spencer, and Richard Goldhor NovaSpeech LLC and [§]Cornell University

ABSTRACT

English cross splicing experiments have shown that speech segments from a variety of speakers and contexts can be substituted for each other in digitized speech without any adverse effect on segmental intelligibility, speaker identity, or naturalness [1, 2]. In many of the sentences tested, more than half the phones in the sentence were spliced in from other sources, yet listeners judged all "surrogate" phones as identical to the originals, and judged the sentences as human-sounding, perceptually coherent, and sounding like the original speaker.

This poster presents our hypotheses about the phonological, acoustic, and perceptual factors that determine which phones can be replaced, and which other phones can serve as surrogates for them. It ends with a discussion of the relevance of these hypotheses for models of speech perception and synthesis, offering new views on some old issues.

INTRODUCTION

The work presented here is part of ongoing work by Hertz and her collaborators over the course of some thirty years in the related areas of multi-language, multi-dialect, and multi-voice speech synthesis by rule; speech perception and timing models; and the phonology/phonetics interface [1-10].

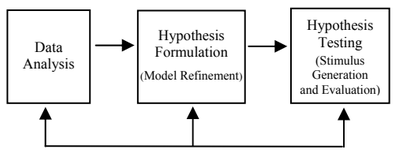
A central quest in all this work has been to identify:

- the role that phonological and perceptual principles play in determining how phonemic and other linguistic distinctions are realized phonetically;
- which information listeners use to identify phonemes;
- which information listeners use to identify speakers (that is, voices)
- what acoustic patterns and perceptual principles are language-universal, language-specific, dialect-specific, speaker-specific, and listener-specific, and
- what factors determine the perception of naturalness in synthetic speech.

In the past two years, we have made extensive use of a particular experimental paradigm—the study of listeners' perception of speech in which selected phonetic segments have been replaced with surrogate speech segments. We have found this technique to yield robust and often surprising results that shed new light on the issues listed above. While we have explored a variety of different types of speech material, we will focus here on the sort we have examined most, namely single-phrase sentences in English with normal declarative intonation and speaking rates.

METHODOLOGY

We constructed hundreds of short- to medium-length sentences in which we tested the perceptual effects of different types of phone-sized waveform substitutions. The specific stimuli examined were selected in stages during an iterative process of hypothesis formulation and testing that has served as the basic methodology for our research in general:



Stimuli were evaluated in informal and formal perceptual tests in which judgments were elicited from listeners concerning naturalness, intelligibility, phoneme identity, speaker identity, speaker characteristics, and other properties [1, 2].

Table 1 describes the types of substitutions made in the stimuli, and highlights one or two selected findings for each category. See **IMPLICATIONS** for a unified summary of the generalizations resulting from the experiments.

SUBSTITUTIONS and FINDINGS

Substitution Dimension original / surrogate	Description of Substitutions Performed and Selected Findings
Speaker e.g. Female / Male Male / Female Male A / Male B Female A / Female B Adult / Child	**Description** Phones from one speaker were replaced by surrogate phones from another speaker. Speakers differed in age, gender, and voice characteristics. **Selected Finding** Most consonants outside of the syllable nucleus can be replaced by surrogates from different speakers (regardless of gender) without affecting speaker identity.
Speech Type Human / Synthetic	**Description** Phones in human speech were replaced by synthesized phones produced with a KLSYN-88 style formant-synthesizer [11]. **Selected Finding** Most consonants, reduced vowels, and many other segments can be replaced by formant-synthesized surrogates without impacting speaker identity.
Spectral Characteristics e.g. [n] / [m] [n] / [s] child [s] / adult [s] dialect1 [l] / dialect2 [l]	**Description** Original phones were replaced by surrogates with markedly different spectral characteristics. Original and surrogate phones differed in manner of articulation, place of articulation, their phonetic context, phoneme affiliation, dialect affiliation, language affiliation, and speaker affiliation. **Selected Finding** Listeners are insensitive to much of the fine-grained spectral detail in consonants, so the same waveform fragment will sound natural in many contexts and with many people's speech.
Duration	**Description** In some stimuli, surrogate durations were adjusted to match the original phones, while in others, the duration of the surrogate segments were used without modification, or were strategically modified to test particular hypotheses. **Selected Findings** In consonants, durations are often more important than spectral values as a cue to segmental identity.
Fundamental Frequency	**Description** In some stimuli, the F0 of phonetically voiced surrogate phones was adjusted to match the F0 of the original utterance, while in others the F0 of the surrogate phone was left unaltered. **Selected Finding** The F0 of phonetically voiced surrogates must be reasonable for the target context in order for the speech to sound coherent.
Variant Pronunciations e.g. [ən] / [ɳ] [tˈ] / [t]	**Description** In some stimuli, phones were replaced with acoustic patterns reflecting acceptable variant pronunciations of the phoneme in question, while in others, phones were replaced with patterns that would never occur. **Selected Finding** Contextually-appropriate variants often sound equally natural, even in cases where the original speaker never produces one of the variants.
Phonation e.g. non-breathy / breathy voiced / devoiced modal voicing glottalized	**Description** Original segments were replaced by surrogates differing in phonation. **Selected Finding** Modally voiced syllable nuclei can often be replaced by contextually permissible glottalized, breathy, or devoiced variants, even from other speakers, without impacting naturalness.

Table 1: Dimensions along which tested surrogate segments have varied.

EXAMPLES

Table 2 describes the voices used in examples in this section. Table 3 shows where each segment in these examples came from. Each surrogate segment was replaced by the comparable segment in the same sentence uttered by a different speaker, except for segments in square brackets, which were taken from the utterances and contexts indicated. Listeners correctly perceive all original phonemes in these sentences, and hear the speaker whose utterance the stressed vowels care from. All sentences sound natural except Sentence 9 (see below).

Speech Type and Speaker	Notes
Human adult females F-SH F-M	F-SH = middle-aged; F-M = young adult, human speaker underlying Mara voice in Speechify 2.0 concatenative TTS system
Human adult males M-JS M-R M-JD M-JFK M-PL	M-JS = middle-aged; M-R = young adult, human speaker underlying Rick voice in Speechify 2.0 concatenative TTS system; M-JFK = John F. Kennedy; M-PL = Peter Ladefoged
Synthetic adult males SM-R SM-F	SM-R = Default male voice (Reed) of ETI-Eloquence 6.0 rule-based TTS system [5]; SM-F = formant synthesizer
Synthetic adult female SF-S	Default female voice (Shelley) of ETI-Eloquence 6.0 rule-based TTS system [5]
Six-year-old female child C-IR	

Table 2: Speakers used in Sample Stimulus Sentences in Table 3

S#	Description	Sounds like:
1: M-R / SM-R	**The expert skier agreed to tee up with the pro golfer** **Human / Synthetic** (8 kHz sampling rate)	M-R
2: F-M / F-S	**It was obvious why the cheetah ate so much food** **Human / Synthetic** (8 kHz sampling rate)	F-M
3: C-IR / F-SH	**The expert skier agreed to tee up with the pro golfer** **Child / Adult** (22.05 kHz sampling rate)	C-IR
4: M-JS / F-SH	**The expert skier agreed to tee up with the pro golfer** **Male / Female** (22.05 kHz sampling rate)	M-JS
5: F-SH / C-IR	**The expert skier agreed to tee up with the pro golfer** **Adult / Child** (22.05 kHz sampling rate)	F-SH
6: F-SH / M-JD	**Monica Naimoo never knew Bonnie's mother** **Female / Male** (22.05 kHz sampling rate)	F-SH
7: F-SH / M-JS	**Monica Naimoo never knew Bonnie's mother** **Female / Male** (22.05 kHz sampling rate)	F-SH
8: F-SH / SM-F	**Mo[n]ica [n]ai[n]oo [n]ever [n]ew Bo[n]ie's [n]other** **Human / Synthetic** (22.05 kHz sampling rate) Note: Same synthetic [n] used for all nasals (see spectrogram in Figure 3) but the original nasal phonemes were still perceived.	F-SH
9: F-SH / F-SH	**[t]o[s]ica [s]ai[s]oo [s]ever [s]ew Bo[s]ie's [s]other** **Female / Female** (22.05 kHz sampling rate) Original nasals (not [s]) were perceived; [s] was perceived as background hiss. See spectrogram in Figure 3.	F-SH
10: M-JFK/ F-SH	**Ask not what your country can do for you** **Male / Female, different dialects** (22.05 kHz sampling rate)	M-JFK
11: M-PL / F-SH	**Peter Ladefoged** **Male / Female, different dialects** (22.05 kHz sampling rate) See spectrogram below.	M-PL

Table 3: Examples of Natural-Sounding and Intelligible Sentences

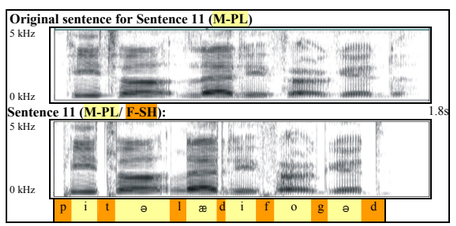


Figure 1. Original and "Surrogated" Versions of *Peter Ladefoged*

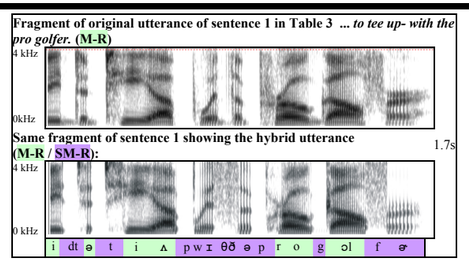


Figure 2. Natural Sounding Human / Synthetic Utterance

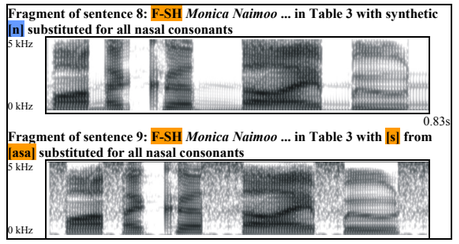


Figure 3. Spectrograms of Fragments of Sentences 8 and 9 (see Table 3)

IMPLICATIONS
Speech Perception

Perhaps the most surprising result of our work is that so many segments, even synthesized ones, sound natural when spliced into contexts very different from the contexts from which they came. Our experiments suggest that in a typical sentence, the *majority* of segments can be synthesized, replaced with surrogate segments from other speakers, or taken from different contexts, without impacting segment intelligibility, speaker identity, or overall naturalness, *provided that certain "rules" are followed*.

We are still learning the nature of these rules, but certain generalizations are emerging:

1. *Speaker Identity*: In order to model a particular speaker, the stressed syllable nuclei must be taken from the speech of that speaker, but it is not clear that any other segments need to be.

***2. Segment Intelligibility*:** As is well known, both formant and timing patterns are important for vowel intelligibility. Vowels also contain important cues to neighboring consonant identity. Like vowels, consonants must have appropriate duration in order to be properly identified. In normal listening environments, however, spectral patterns in consonants seem to be most relevant for high-intensity consonants such as [s]. Even for these consonants, the patterns only have to be *grossly appropriate* (e.g., [s] must have intense, relatively high-frequency noise). This sort of "ballpark perception" is evidenced by the fact that the same nasal or non-nasal obstruent can serve as a natural-sounding and voice-appropriate surrogate for a wide range of speakers, including adults and children, males and females, so long as its duration, amplitude, and—in some cases—F0, fit the phonological and phonetic context.

3. *Naturalness*: Our experiments indicate that for natural-sounding speech, more segment types than just nasal and non-nasal obstruents can serve as surrogates. We have successfully replaced non-nuclear sonorants such as [l] and [w], reduced vowels, syllabic nasals and sonorants, and even phrase-final vowels (see, for example Sentences 1-2, 6-8, 10-11). In longer sonorants, formant values must be generally appropriate for the speaker's dialect and vocal tract size, though these values do not have to be values that the speaker would ever produce. Since actual formant values may be important for naturalness in these cases, formant synthesis, rather than natural-speech waveform concatenation, offers the greatest potential for producing satisfactory non-consonantal surrogates.

More generally, our results strongly indicate that:

Durational patterns in consonants play a dominant role in perception: In natural speech, duration is often used to distinguish phonemes even when those phonemes are realized with markedly different spectral structures, such as [s] and [d]. If one shortens the duration of the [s] in [asa] appropriately, one will hear [d] (albeit with background noise). (See also Sentence 9). Duration often comes to the rescue when other potential cues, like formant transitions, are ambiguous.
One implication of the dominance of duration over spectral detail is that one can often use a single phone from a single speaker as a surrogate for many phones across many speakers. For example, [n] can be used for both [n] and [m] in a wide range of speakers, so long as the F0 and durations are appropriately adjusted for the context.

Consonant clusters are perceptual unit:
Consonant clusters often function as single durational and perception units, and must thus be treated as unitary surrogate units. For example, if a segment is lenited from a cluster, other segments will often compensate durationally in such a way that the total cluster duration remains relatively stable. Through this compensation mechanism, we may still hear a phoneme that is not segmentally present—for example, /nts/ when the signal only contains [ns].

There are many perceptual universals: We continue to find strong evidence for acoustic and perceptual language universals. One of many examples is the directions, durations, and relative durational stability of formant transitions between segments of particular places and manners of articulation [5, 10]. The acceptability of segments from other languages as surrogates in English suggests that our experimental paradigm will yield new insights into universal structures and processes across languages

Humans use selective attention and processing:
Our work suggests that listeners attend to different parts of the speech signal selectively for different purposes, such as determining speaker identity and extracting abstract linguistic units like phonemes. While the perception of stressed vowels seem to involve a memory of what a particular speaker sounds like, other research we are conducting suggests that consonant "chunks" are processed in more knowledge-based ways for purposes of speech parsing, a result consistent with the perceptually robust nature of the consonants manifested in the surrogate experiments.

There is a great deal of invariance in speech:
It is this invariance that makes it possible to substitute whole groups of segments from the speech of one speaker into that of another without affecting the perception of timing, segmental identity, or naturalness. Much of the invariance is in durational patterns, which, as mentioned above, are strong perceptual cues to segment identity. Interestingly, our experiments suggest that speech exhibits a high degree of variability precisely and only in those cases where the variability is not perceptually salient.

Speech Synthesis

State-of-the-art formant synthesis systems can produce highly-natural segments of every type except stressed vowels: The segments that we found to sound consistently non-human in *continuous speech* (and which we were generally unable to modify to sound human) are modally-voiced stressed vowels—the very segments that seem to be most important for characterizing particular voices.

Hybrid synthesis is a promising technology:
Our results support the viability of a hybrid synthesis system in which a natural speech database is used to supply stressed vowels, and formant-based rules are used to synthesize all other segments [2]. The results suggest that such a system can achieve natural-sounding and intonationally-appropriate speech that mimics a particular speaker in a much more flexible way, and with far less data storage and data preparation requirements, than can any current synthesis technology.

[1] Hertz, S. R. , Spencer, I., Church, T. F., and Goldhor, R. (2004) Perceptual consequences of nasal surrogates in English: implications for speech synthesis, *Proc. 147th Meeting of the Acoust. Soc. America*, New York.
[2] Hertz, S.R. (2002) Integration of rule-based formant synthesis and waveform concatenation: a hybrid approach to text-to-speech synthesis, *Proceedings IEEE 2002 Workshop On Speech Synthesis*, Santa Monica, CA.
[3] http://www.mindspring.com/~ssshp/ssshp_cd/ss_corn.htm (Smithsonian speech synthesis history).
[4] http://www.mindspring.com/~ssshp/ssshp_cd/ss_eloq.htm (Smithsonian speech synthesis history)
[5] Hertz, S.R., Younes, R.J., and Zinovieva, N. (1999) Language-universal and language-specific components in the multi-language ETI-Eloquence text-to-speech system. *Proc. 14th Int. Cong. Phonetic Sciences*, 2283-2286.
[6] Clements, G.N., Hertz, S.R. (1996) An integrated approach to phonology and phonetics, in J. Durand and B. Laks (eds.), *Current Trends in Phonology: Models and Methods*, CNRS, Paris X and University of Salford Publications.
[7] Hertz, S. R. and Huffman, M. K. (1992) A nucleus-based timing model applied to multi-dialect speech synthesis by rule, *Proc. International Conference on Spoken Language Processing* **2**, 1171-1174.
[8] Clements, G. N. and Hertz, S. R. (1991) Nonlinear phonology and acoustic interpretation, *Proceedings of the XIIth International Congress of Phonetic Sciences* **1**, Aix en Provence (France), 364-373.
[9] Hertz, S. R. (1991) Streams, phones, and transitions: toward a phonological and phonetic model of formant timing, *J. of Phonetics* **19**, *Special Issue on Speech Synthesis and Phonetics*, ed. by R. Carlson.
[10] Hertz, S. R. (1990) The Delta programming language: an integrated approach to non-linear phonology, phonetics, and speech synthesis, in J. Kingston and M. Beckman (eds.), *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, Cambridge University Press.
[11] Klatt, D. H. and Klatt, L. C. (1990). Analysis, Synthesis and Perception of Voice Quality Variations among Female and Male Talkers, *J. Acoust. Soc. America* **87**, 820-857.

Acknowledgements
The work was supported in part by NIH grant R43 DC006761-01 to NovaSpeech LLC. The views are those of the authors and do not necessarily reflect those of the agency.

[1]This poster is a slightly revised version of one presented at the July 2004 conference *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, MIT, Cambridge, MA.