

# WHEN CAN SEGMENTS SERVE AS SURROGATES?<sup>1</sup>

Susan R. Hertz<sup>§</sup>, Isaac C. Spencer, and Richard Goldhor NovaSpeech LLC and <sup>§</sup>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:

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

## SUBSTITUTIONS and FINDINGS

Substitution Dimension original / surrogate	Description of Substitutions Performed and Selected Findings
<b>Speaker</b> e.g. Female / Male Male / Female Male A / Male B Female A / Female B Adult / Child	<b>Description</b> Phones from one speaker were replaced by surrogate phones from another speaker. Speakers differed in age, gender, and voice characteristics.  <b>Selected Finding</b> Most consonants outside of the syllable nucleus can be replaced by surrogates from different speakers (regardless of gender) without affecting speaker identity.
<b>Speech Type</b>  Human / Synthetic	<b>Description</b> Phones in human speech were replaced by synthesized phones produced with a KLSYN-88 style formant-synthesizer [11].  <b>Selected Finding</b> Most consonants, reduced vowels, and many other segments can be replaced by formant-synthesized surrogates without impacting speaker identity.
<b>Spectral Characteristics</b>  e.g. [n] / [m] [n] / [s] child [s] / adult [s] dialect1 [I] / dialect2 [I]	<b>Description</b> 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.  <b>Selected Finding</b> 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.
<b>Duration</b>	<b>Description</b> 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.  <b>Selected Findings</b> In consonants, durations are often more important than spectral values as a cue to segmental identity.
<b>Fundamental Frequency</b>	<b>Description</b> 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.  <b>Selected Finding</b> The F0 of phonetically voiced surrogates must be reasonable for the target context in order for the speech to sound coherent.
<b>Variant Pronunciations</b>  e.g. [ː n] / [nβ] [t^] / [t]	<b>Description</b> 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.  <b>Selected Finding</b> Contextually-appropriate variants often sound equally natural, even in cases where the original speaker never produces one of the variants.
<b>Phonation</b>  e.g. non-breathy / breathy voiced / devoiced modal voicing glottalized	<b>Description</b> Original segments were replaced by surrogates differing in phonation.  <b>Selected Finding</b> 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]a[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]a[s]oo [s]lever [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

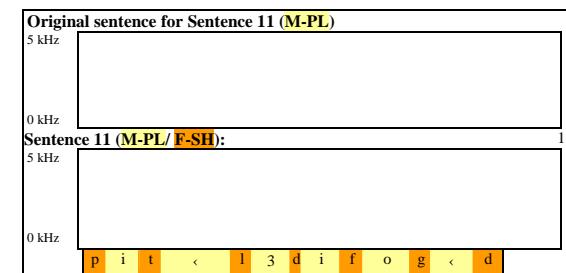


Figure 1. Original and “Surrogated” Versions of Peter Ladefoged

Fragment of original utterance of sentence 1 in Table 3 ... to tee up- with the pro golfer. (M-R)



Same fragment of sentence 1 showing the hybrid utterance (M-R / SM-R):

