

## 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-phoneme speech synthesis by rule, speech perception and timing models, and the phonology-phonetics interface [1-10].

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

- the role that phonological and perceptual principles play in determining the perceptual distance between natural and/or synthetic (real or virtual) phonemes.
- which information listeners use to identify phonemes (that is, vowels, consonants, and syllable structures).
- what acoustic patterns and perceptual principles are language-dependent, language-specific, dialect-specific, speaker-specific, and/or larynx-specific.
- what factors determine the perception of naturalness in various speech situations.

In the past two years, we have made extensive use of one particular experimental paradigm—the study of listeners' perception of speech in which one or more phonetic features of the target speaker have been replaced with surrogate speech segments. We have found this technique to be robust and often surprisingly unambiguous that had not been so in the past.

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-phoneme utterances in English with normal normal articulatory movements, but with synthesized or synthesized voices.

**METHODS**

We constructed hundreds of short-to-medium length sentences in which a single phonetic feature of the target speaker has been replaced with surrogate speech segments. The specific stimuli examined were selected in each case during an iterative process of hypothesis formulation and testing that has served as the basic methodology for our research in general:

- **Hypothesis Formulation**: Model Refinement
- **Hypothesis Testing**: Model Validation
- **Data Analysis**: Model Evaluation

**STIMULI**

Stimuli were evaluated in informal and formal perceptual tests in which judgments were elicited from listeners concerning various speech attributes like identity, speaker characteristics, and other properties [1, 2].

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

<table>
<thead>
<tr>
<th>Substitution type</th>
<th>Description of Substitutions Performed and Selected Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Speakers</strong></td>
<td>Phrases from one natural speaker were replaced by surrogate phrases from another speaker: Discriminated by age.</td>
</tr>
<tr>
<td><strong>Selected Findings</strong></td>
<td>Most consonant replacements were made by replacing consonants from different speakers with consonants from the same speaker.</td>
</tr>
<tr>
<td><strong>Phoneme Clusters</strong></td>
<td>The effect of replacing individual segments varied depending on the context of the segment.</td>
</tr>
</tbody>
</table>

**Instrumentation**

- **Microphones**: Two different types of microphones were used: 1) A Shure SM58 and 2) A Sennheiser MD431.
- **Sampling Rate**: All recordings were made at a 44.1 kHz sampling rate.
- **Equalization**: All recordings were equalized to a standard level.
- **Normalization**: All recordings were normalized to a standard level.

**RESULTS**

The expert skier agreed to tee up with the pro golfer in the professional match. The expert skier was able to win the match, while the pro golfer was able to lose it. The expert skier was able to win the match, while the pro golfer was able to lose it. We are still learning the nature of these rules, but certain generalizations are:

- **1. Speaker Identity**: For gender identity, the stressed nasal phonemes were replaced with corresponding segments from the original nasal phonemes were still perceived. \[\text{Original nasal phonemes were still perceived.} \]
- **2. Segment Identity**: As is well known, both formant and timing information are used in order to perceive identity. The most important source of timing information is the length of the nasal phoneme. \[\text{The most important source of timing information is the length of the nasal phoneme.} \]
- **3. Phoneme Clusters**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **4. Acoustic Characteristics**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **5. Spectral Characteristics**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **6. Duration**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **7. Variants**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **8. Phonology**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]

**DISCUSSION**

- **Hypothesis Testing**: Model Validation
- **Data Analysis**: Model Evaluation

**CONCLUSIONS**

- **1. Speaker Identity**: For gender identity, the stressed nasal phonemes were replaced with corresponding segments from the original nasal phonemes were still perceived. \[\text{Original nasal phonemes were still perceived.} \]
- **2. Segment Identity**: As is well known, both formant and timing information are used in order to perceive identity. The most important source of timing information is the length of the nasal phoneme. \[\text{The most important source of timing information is the length of the nasal phoneme.} \]
- **3. Phoneme Clusters**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **4. Acoustic Characteristics**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **5. Spectral Characteristics**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **6. Duration**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **7. Variants**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]
- **8. Phonology**: Nasal phonemes were perceived differently depending on the context of the segment. \[\text{Nasal phonemes were perceived differently depending on the context of the segment.} \]

**REFERENCES**

[5] Acknowledgements

The work was supported in part by NIH grant R01-HL65761-06 to Eloquence, Inc. and by a variety of other grants and foundations, including the Canadian Institute for Speech and Hearing Research, the National Science Foundation, the National Institutes of Health, and the National Science Foundation. The authors do not necessarily reflect those of the agency.