The Use of Random Spliced Speech for the Recognition of Familiar Voices

Ricardo M. de Figueiredo

Department of Legal Medicine, State University of Campinas, Campinas, Sao Paulo, Brazil, Cx. Postal 6153, CEP 13084-970, ricfig@turing.unicamp.br

Abstract: Various techniques for experimental speech signal manipulation to test recognition of familiar voices have been used, but random-spliced-speech (RSS), which offers the advantage of more natural-sounding speech and a decreased risk of the creation of artifacts interfering with perception has not been tried. The present paper describes an experiment in which the recognition of familiar voices submitted to random splicing was investigated.

INTRODUCTION

Various masking procedures have already been used in experiments involving perceptual recognition of familiar voices, including selective filtering (Pollack et al., 1954; Clarke and Becker, 1968; Abberton and Fourcin, 1978), analysis and synthesis techniques (Brown, 1981; van Dommelen, 1990; Kuwabara and Takagi, 1991), backward speech (Bricker and Pruzansky, 1966; van Lancker et al., 1985a), and compressed/expanded speech (van Lancker et al, 1985a). Such manipulations were generally designed to emphasize or neutralize specific acoustic cues. Masking procedures, however, tend to create a kind of speech that sounds much too artificial and/or degraded. An alternative technique is random-spliced-speech (RSS), which compares favorably with other techniques. While sounding more natural, it truly eliminates original intonation contours, yet retains most of the cues to voice quality, as well as F0 level and range.

METHODS

Eleven male native speakers of Brazilian Portuguese read a text of 126 words. Four of them were speakers of the Northeastern Brazilian dialect, while the others speak another dialect. Three of the speakers from the Northeast were selected as the target voices to be recognized by listeners who were already familiar with their voices.

All speech samples were recorded in a room with no special acoustic treatment, but with a negligible background noise level, using a Tascam DA-30 DAT recorder and a Realistic 33984-C microphone. The taped signals were digitized by 16-bit ADC at a sampling rate of 10 Khz. The original speech was cut into segments of 250 ms, and these were randomly recombined, with adjacent segments overlapping by 10 ms. The overlapping portions were linearly attenuated to zero amplitude to avoid transients. For each speaker (n=11), eight RSS samples of 3.5 s were selected, a total of 88 different stimuli (8 x 11).

The listeners (n=34) were divided into three approximately equal groups, each group being familiar with one and only one of the target voices. The stimuli were arranged in random order, with each one being presented twice, separated by a single second pause. On an answer sheet, listeners were instructed to indicate for each and every stimulus whether it was or was not the specific familiar target voice. To train listeners for the task, a pre-test was conducted using the voices of celebrities.

RESULTS AND DISCUSSION

In this experiment, only two types of error - Type I and Type II - are possible: false acceptance (erroneous recognition of a stimulus as being the target voice) and false rejection (failing to recognize the target voice), respectively. Although familiar target voices were almost always recognized, the most frequent misidentifications were of other voices as being the same as this voice (93% of the errors committed).

The expected number of correct identifications by chance alone is determined by the equation $E = kmn$, where $k$ is the total for affirmative answers (correct + Type I errors), $m$ is the number of stimuli with the target voice, and $n$ is the total
number of stimdi, Figure 1 shows the number of correct recognitions in relation to $k$ and $E$ for each listener. It can be seen that although performance may vary considerably, all listeners recognized the target voice well above the level of chance.

![Figure 1](image)

**FIGURE 1.** Correct recognition (thick line), affirmative recognition (dotted line), and chance level (thin line). The listeners were arranged along the horizontal axis in order of increasing affirmative answers (correct + Type I error).

A statistical analysis of misidentifications showed that if the difference between the average $F_0$ of two speakers is less than 2.8 semitones, there is a significant increase in the confusion between the two speakers. The proximity of the Long Term Average Spectra (LTAS) also influenced misidentifications, with voices revealing greater correlation between LTAS tending to be more confused. However, there is also a significant interaction in relation to the behavior of $F_0$, since mistakes related to proximity of LTAS only increased when the difference between the average $F_0$ of the two speakers was less than 2.8 semitones. Speakers of the same dialect were no more likely to be identified erroneously than those of other dialects, suggesting that dialectal cues no longer remain in RSS.

Degree of familiarity with the target voice was evaluated by asking each listener approximately how many hours per week he had contact with the target speaker. A significant negative correlation was found between the degree of familiarity expressed in contact hours per week and the number of errors ($r = -0.69$, $p < 0.005$).

Personal ability also seems to play a role in the process of recognition, as a significant positive relation ($r=0.74$, $p < 0.005$) was found between the number of errors on the pre-test (voices of famous people) and the number of errors in the final experiment.

**SUMMARY**

In summary, familiar target voices are almost always recognized, even when RSS is used, although performance varies from one individual listener to another, depending on degree of familiarity and intrinsic ability; misidentifications are apparently unrelated to dialect similarity, tending to be the false recognition of other voices as being the same as the target, especially when average $F_0$s and LTAS are very close.

**REFERENCES**