PERCEPTUAL RECOGNITION OF CELEBRITY VOICES USING RANDOM SPLICED SPEECH

Ricardo Molina de Figueiredo
State University of Campinas (UNICAMP), Brazil

ABSTRACT

The present paper describes an experiment investigating in the perceptual recognition of the voices of celebrities. Stimuli of 3.5 seconds in length were presented in the form of random spliced speech, a masking technique which totally destroys the intelligibility of speech, while maintaining various factors important for recognition, such as overall voice quality, F0 level and F0 range. The results reveal a recognition index well above chance level.

1. INTRODUCTION

Familiar voices are well recognized perceptually, even if the stimulus is presented in the form of Random Spliced Speech (RSS) [1]. Experiments with familiar voices, however, are rather problematic, since degree of familiarity is difficult to evaluate. Moreover, the number of listeners who are thoroughly familiar with a specific voice is generally limited.

Evaluating familiarity by asking each listener approximately how many hours per week he had contact with the target speaker, [1] found a statistically significant negative correlation between the degree of familiarity and the number of errors. Although this correlation was to be expected, with greater familiarity linked to fewer errors, certain doubts remain as to the reliability of self-evaluation to determine degree of familiarity. Although an attempt was made to minimize the variation between listeners by forming groups of colleagues who work with the target speaker, this cannot guarantee that some other factor such as, for example, degree of personal involvement (a qualitative dimension not directly reflected in "number of hours of weekly contact") will not influence performance.

Another option to ensure familiarity is to use the voices of celebrities in such perceptual tests of recognition. This offers certain advantages and avoids some of the problems. First of all, there is no natural limit as to the number of listeners used in each test, since huge numbers of individuals regularly hear celebrities on television. It also seems reasonable to assume that the degree of familiarity becomes less critical, since a more impersonal relationship between listener and target speaker can be assumed.

Another advantage involves the possibility of repeating the same experiment with other groups of listeners who are equally representative, and equally familiar with the target voice. This is especially important when it is considered that the performance on repeated tests of perceptual recognition of voices (whether or not they are familiar) can be strongly influenced by training. In other words, a second test with the same target voice presents fewer difficulties for a listener, which contributes to an undesirable response bias.

In conclusion, the use of the voices of celebrities ameliorates some of the complications encountered when evaluating voice recognition, making it possible that a) many more listeners can be evaluated with a single stimulus tape, b) familiarity with the target voice can be assumed to be relatively homogeneous for the listeners, and c) experiments (or variations of them) can be replicated, thus minimizing the risk of training effects in listener performance.

2.METHODS

The speech of three Brazilian television celebrities (an actor frequently appearing in popular daily soap operas, a television news commentator, and the host of a talk show) was taped directly from the audio outlet of a standard television set and digitalized at 16 Khz, 16 bit ADC. The speech samples were then processed to obtain RSS. During this procedure, the original speech is cut into 250-ms segments, and these segments are randomly combined, with adjacent segments subjected to 10-ms overlapping. The overlapping segments were then linearly attenuated to zero amplitude to avoid transients, which might interfere with perception.

Numerous 3.5-second RSS samples were created for each target speaker, and eight of them were selected. The RSS samples of ten other speakers were produced in the same way, thus creating test tapes with 88 randomly ordered stimuli, eight from the target speaker, and 80 from the other speakers (8 x 10).

The test tape was designed to present each stimulus twice during a one-second interval, with a four-second interval left after each stimulus to provide time for listener response. Twenty-eight listeners (14 men; 14 women) filled out an answer sheet for each of the voices, indicating whether it was or was not the target voice (forced answers). The listeners did not know how many times the target voice would occur on each test tape. Prior to the test itself, all listeners took a pre-test to ensure that the task was clear, with a different target voice used to avoid training effects.

The test stimuli were presented through good-quality commercial headphones in rooms with a very low noise level, although no special acoustic treatment was considered necessary. Separate sessions were conducted for each target voice, with approximately a week between them. Six sample groups of listeners were formed (n=5, 5, 5, 5, 4, 4) so that each target voice was tested in each of the possible sequences (3!= 6) to avoid eventual order effects for specific voices.

3.RESULTS

The performance of the listeners in this type of test can be evaluated more objectively by comparing it to the number of correct recognitions expected due to chance. This expectancy index E can be calculated from the equation $E=km/n$, where k is the total number of 'yes' answers, m is the number of stimuli of a given target voice, and n is the total number of stimuli presented for the test. Since m and n

are constants (8 and 80, respectively, in this experiment), the level of chance depends only on the number of 'yes' answers, maintaining a linear relation with *E*.

Figure 1 illustrates the results obtained. The shaded polygon reveals the limits of observed correct recognitions as a function of observed 'yes' answers, while the slanted line shows the number of correct answers expected *E* for each number of 'yes' answers. It is clear that, in all cases, the index of correct recognition is well above chance level.

In this experiment, only two types of errors were possible: false acceptance (FA, i.e., erroneous recognition of a stimulus as being the target voice) and false rejection (FR, i.e., failure to recognize the target voice). The bar graph in Figure 2 shows the average number of each of these types of error, together with the expected level for chance occurrence. Both FA and FR errors are much less frequent than chance level, which is not surprising, given the fact that the average index of correct answers is well above chance level. However, it can be seen that the latter are much less frequent. In other words, listeners are more likely to accept a non-target voice as being a target voice than to fail to recognize a target voice.

An analysis of variance (ANOVA) was used to test the influence of various factors on the weighted index of correct answers (i.e., number of correct answers remaining after subtracting the corresponding expected number of 'yes' answers). The performance of individual listeners revealed statistically significant variation (F=9.78; p<0.001), which indicates that some individuals are more accurate in their perceptions than others. Sex also exerts a significant effect, with women performing slightly better than men (male = 5.53 hits; female = 6.32 hits; F=17.52; p<0.001). A marginal section effect was also observed (F=3.92; p = 0.05), indicating an improvement in performance from the first to the third section, presumably due to practice (S1=5.52 hits; S2=5.78 hits; S3=6.41 hits). The specific target voice, however, had no significant effect, i.e., none of the voices of the celebrities used in this experiment was more frequently recognized than any other.

4.CONCLUSION

The results obtained here prove that listeners are capable of recognizing the voices of various celebrities presented in the form of Random Spliced Speech well above chance level, confirming results obtained for familiar voices in [1]. The fact that voices can be recognized in the form of RSS has made this technique potentially useful in certain forensic situations. By destroying the intelligibility and possibly some emotional connotations of the content of the original speech sample, RSS stimuli should minimize certain subjective aspects involved in the voice identification of witnesses.

Various aspects are left to be explored in relation to perceptual recognition using RSS, however. One of these is the reaction time of the listener. The measurement of the time lag between stimulus presentation and subject response may lead to the development of an "uncertainty index" and be a step towards the refinement of the evaluation process of listener responses, possibly leading to the proposal of an alternative to the dichotomy of forced yes/no tests.

Another aspect involves the response bias in repeated tests. As [2] have suggested, a simple statistical model cannot account for repeated perceptual recognition tasks, since such tests cannot be considered as independent events. In other words, it is difficult to evaluate whether a listener is recognizing a sample of speech based on some internal reference, or whether it is being compared to a previous sample in the test itself, which has been assumed to be that of the target speaker.

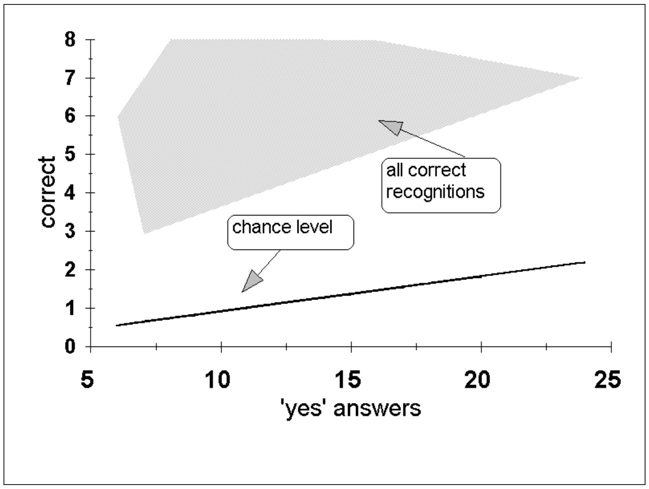

Figure 1. ‘Yes’ answers *vs*. correct answers. The shaded polygon limits the area of all correct recognitions. The line shows the chance level, i.e., the expected number of hits for each number of ‘yes’ answers.

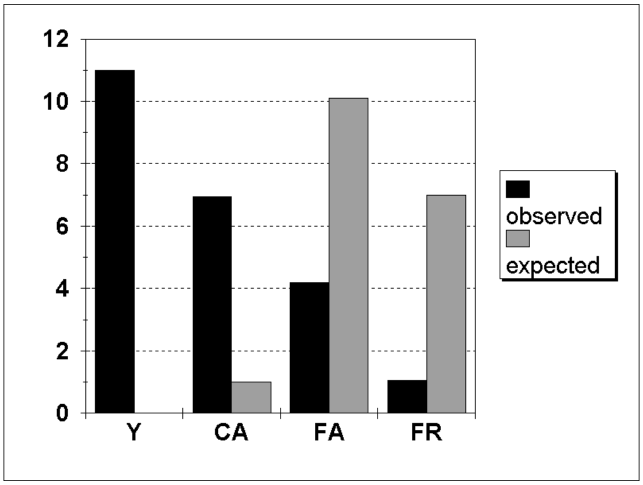

Figure 2. Average number of ‘yes’ answers (Y), correct answers (CA), false acceptance errors (FA), and false rejection errors (FR). Black bars are observed values, and gray bars are expected values by chance alone.

REFERENCES

[1] Figueiredo, R.M. 1998 The use of random spliced speech for the recognition of familiar voices”, *Proceedings of the 135th Meeting of the Acoustical Society of America*, 1299-1300

[2] Broeders, A.P.A. and Rietveld, A.C.M. 1995 Speaker Identification by earwitnesses, in A. Braun and J-P. Köster (eds) *Studies in Forensic Phonetics*, Trier: Wissenschaftlicher Verlag, 64: 24-40