Relative Importance of Parameters in Voice Similarity Judgment

R. Brown
Singapore

The field of auditory speaker recognition is concerned with the ability of human listeners to recognise a speaker's identity from hearing a sample of his speech. It involves a pattern-matching technique; on hearing the sample, the listener abstracts a representation of the voice which he then compares with an internalised reference pattern. Research in the field has concentrated on specifying those acoustic features which compose such voice patterns. Many experimenters have manipulated one feature in isolation, or isolated the glottal or vocal-tract contributions to voices (laryngograph, vocoder, inverse filtering, whispering, using an electrical larynx, etc.). However, results from such experiments indicate that each of the features investigated in isolation contributes something to speaker recognisability. Of greater relevance, therefore, to not only experimental, but also everyday speaker recognition is a statement of the *relative* importance of features.

The task in the present experiments is one not strictly of speaker recognition, but of voice similarity judgment, on the principle that the more similar a pair of voices are judged to be, the more difficult they will be to differentiate in a speaker recognition experiment, and vice versa. Synthetic voices were used, produced on a PAT synthesiser (Anthony and Lawrence, 1962). Stimulus samples consisted of various combinations of high and low values for the eight parameters below. The control sample, with which stimulus samples were paired for comparison, contained mid values for all eight parameters. With the exception of parameters 3, 6 and 7 below, these mid values were taken from a live utterance by the author.

1. Formant (F) range.
   High: approx. 30% increase in control value
   Low: approx. 30% decrease in control value
2. F mean.
   High: 15% increase in control value
   Low: 15% decrease in controle value
3. F bandwidth.
   High: 150 Hz. Control: 100z.
   Low: 50 Hz.
4. Fundamental frequency (F_0) mean.
   High: 20% increase in control value
   Low: 20% decrease in control value

5. F_0 range.
   High: approx. 45% increase in control value
   Low: approx. 45% decrease in control value
6. Larynx amplitude mean.
   Agreed auditory categorisations of loud, moderate and quiet (owing to lack of instrumentation)
7. Whisperiness.
   Agreed auditory categorisations of extreme, moderate and slight whisperiness
8. Tempo mean.
   High: 10% increase in control value
   Low: 10% decrease in control value

A homogeneous set of listeners were required to judge the similarity of pairs of voices (the control followed by a stimulus) on a 100-point scale ranging from SIMILAR (0) to DIFFERENT (100). A short-term memory task was set between the randomised trial presentations. Two replicates of a one-quarter replicate factorial design were employed, presenting 8 listeners with 16 trials each. A second experiment was carried out employing a full factorial design with 16 listeners, 16 trials per listener and the first four of the above factors, selected on the basis of the results of the first experiment. The results of the second experiment are therefore more reliable than those of the first.

Table I indicates the main effects and the 6 most important second-order interactions of the factors. These are expressed as shifts along the 100-point

Table I. Main effects and second-order interactions in the 2 voice similarity judgment experiments

	Experiment 1	Experiment 2
Main effects		
F range	1.87	−1.65
F mean	5.88**	−7.93**
F bandwidth	−4.48**	−10.07**
F_0 mean	−3.52*	−4.96**
F_0 range	−2.23	
Amplitude mean	1.76	
Whisperiness	0.54	
Tempo mean	−6.40**	
Second-order interactions		
F range/F mean	−4.99**	5.86**
F range/F bandwidth	−1.85	−5.54**
F range/F_0 mean	−1.46	1.22
F mean/F bandwidth	−0.37	−1.30
F mean/F_0 mean	0.80	−2.48*
F bandwidth/F_0 mean	0.13	−1.77

* Significant, $p < 0.05$.
** Significant, $p < 0.01$.

response scale (positively towards the DIFFERENT end, negatively towards the SIMILAR end). Main effects represent half the average difference in response between samples containing the factor at the high level against the low. Second-order interactions indicate the effect of having both factors at the same (high or low) level.

There are three main conclusions:

1. There is justification for the adoption of a design implying a linear model whereby a listener's response for a particular factorial combination is expressed as the sum of the mean response for that listener, the values of the appropriate main effects and interactions and an error factor. Although listeners differed in the average level of performance, their reactions to changes in the factors did not differ significantly.
2. F mean, F bandwidth and F_0 mean were consistently found to be significant. The reverse in polarity between the two experiments for F mean and for F mean/F range is, however, worrying.
3. Tempo mean, found to be significant in Experiment 1, deserves further investigation as a speaker-characterising feature.

Reference

Anthony, J. and Lawrence, W. (1962). A resonance analogue speech synthesizer. *Proceedings of the 4th International Congress on Acoustics, Copenhagen.* Paper G43.