Discriminating Same or Different speech: human vs machine

Greg Kochanski, Christina Orphanidou and Burton S. Rosner

Phonetics Laboratory, The University of Oxford, Oxford, UK `greg.kochanski@phon.ox.ac.uk`

This paper is submitted to Interspeech 2007, Aug 27-31, 2007, Antwerp, Belgium. It may be found at `http://kochanski.org/gpk/papers/2007/ICSLP2007.pdf`

Abstract

We compare the performance of a Bayesian classifier against humans in a same/different speech perception task. The classifier is trained on different sizes of speech segments to separate sounds into perceptually equivalent and nonequivalent categories. We test the classifier with pairs of speech sounds of various sizes and compare its performance with perceptual listening tests on native speakers of Southern British English as well as Greek native speakers who speak English as a second language. We find that the performance of the classifier is comparable to that of human subjects and that the largest-sized speech segments produce the best performance overall. We also find that an edge detector in the 60-800 Hz band and an autocorrelation-based measure of voicing are very important to the performance of the classifier.

Index Terms: phonological difference machine learning

1. Introduction

There is yet no complete answer of what determines whether two sounds are perceptually equivalent or not. This question is relevant to auditory speech perception and to the development of speech models for applications. Many studies have attempted to shed some light to the topic. Pitch perception has been associated with frequency [8] [9], temporal intensity [11] as well as the spectral structure of the sound [10]; loudness detection [16, 18] and other psychophysical properties of speech have been studied [13] with variable results. Measures of perceptual difference have also been developed to evaluate speech coders ([17, 15] and references therein).

In the current study we model speech with a spectral acoustic description vector and build a Bayesian classifier that is tuned to respond to perceptually relevant differences. Unlike most previous studies on auditory speech perception, the stimuli we use are not phonemes, syllables or words. Instead we use small windows of speech of different sizes randomly picked from a phonetically rich speech corpus. After training, we test the classifier by feeding it with pairs of sounds to determine whether they are "matched" or "unmatched". We then compare its performance with perceptual decision tests of native speakers of Southern British English as well as Greek native speakers that speak English as a second language.

2. Experimental methods

2.1. Producing the Stimuli

The stimulus production began with eleven native speakers of the Southern British English dialect. Each subject read out a list of approximately 250 phonetically rich sentences. Once segmented, an acoustic description vector (see §3.1) was computed for the audio at 5 ms intervals. We chose pairs of utterances that were read from the same text (93% were from different speakers), and time-aligned the pairs using the dynamic time-warping algorithm from [19]. From these time-aligned pairs, we extracted 22088 pairs of speech samples; half were from matched moments in the pair, and half were mis-matched by at least 250 ms. The speech samples were 22, 45, or 100 ms long, and were extracted using a $\cos^2$ window to eliminate clicks. The centers of the two sounds were separated by 0.3 s, 1.3 s, or 2.3 s of silence[1]. An example is shown in the top panel of Figure 1.

2.2. Perceptual Tests

We used two groups of linguistically naive subjects. Ten were native speakers of Standard Southern British English. Eight were native speakers of Modern Greek who lived in Greece until at least adulthood, speak English as a second language but have not lived in an English speaking country for more than 6 years. (Qualitatively, they are fluent in English but have noticeable accents.) All were undergraduate and postgraduate students at Oxford University. Subjects were seated at a computer in a quiet room and wore a set of headphones; the experimenter spoke to each subject in their L1.

In the task, subjects were presented with pairs of sounds by a computer program and were asked to indicate whether they thought the sounds came from the same part of the same word or if it was different by pressing a button. Before the experiment started subjects were presented with 8 pairs of matched and unmatched sounds, with correct answers provided, to familiarise them. A to-

[1] We found that there was no statistically significant effect of the length of the silence between pairs. Thus, we lumped the different silence intervals together in the analysis.

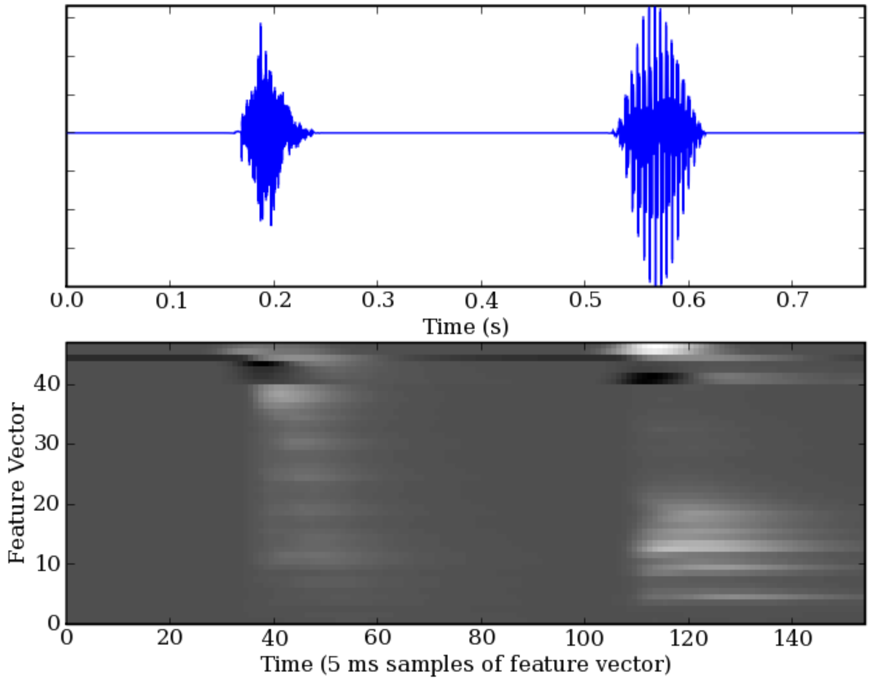

Figure 1: The top panel shows the audio signal for a stimulus. The bottom panel shows the audio description vectors as a grey scale. Vectors run vertically with the spectrum components at the bottom and voicing estimates on top.

tal of 234 stimuli was presented to each subject, randomly selected from a pool of 3456. Subjects reported that they were confident that they had performed well.

3. Analysis

We compare a machine classification of the signals to two groups of human judgements. The machine operates on the signals that were presented to the human subjects as described in (§2.1).

3.1. Acoustic Description Vector

Training the classifier begins by computing a feature vector. It is computed by feeding the speech signal into a filter bank spanning 60–6000 Hz with cochlear filters (taken from [1]) spaced every 0.71 erb. The filtered outputs are half-wave rectified. Then, the signal is divided by the the iso-loudness contour of human hearing at 70 phons at the filter's center frequency. We define $\rho(t,\omega)$ to be the 2/3 power of the above. (This power law approximates the perceptual loudness response.)

Then, we approximate the modulation transfer function of the human auditory system, by smoothing $\rho(t,\omega)$ in the time domain with a single-pole causal low-pass filter, whose time constant is from [2], yielding $r(t,\omega)$. (This is analogous to the "perceptual spectrum" from [6].)

To make the result independent of the recording sensitivity, we normalise r and smooth to compute a normalized perceptual spectrum:

$$R(t,\omega) = \kappa_1 \times r(t,\omega)/(\kappa_2 \times \bar{r}(t) + \alpha\bar{\bar{r}}), \qquad (1)$$

where $\bar{r}(t)$ is the average of $r(t,\omega)$ over all frequency channels and thus proportional to the specific loudness. Also, $\bar{\bar{r}} = \sum_t \bar{r}(t)^2 / \sum_t \bar{r}(t)$, which is approximately proportional to the average loudness of the utterance [18]. Informal experiments on another data set led us to use $\alpha = 0.75$. The "$\times$" operator means time-domain convolution, and κ_1 and κ_2 are 15 ms wide boxcar functions.

The feature vector is composed of 47 components:

1. $R(t,\omega)$, as above (40 components).

2. A part designed to detect changes in the spectrum (edge detectors). Here, we filter into four broad frequency bands, set κ_1 to differentiate on a time scale of 60 ms and κ_2 to smooth over a corresponding window.

3. One component is a spectral entropy measure, inspired by [3], and computed from $r(t,\omega)$.

4. Two final components are indicators of voicing, inspired by [4]. Autocorrelations of $\rho(t,\omega)$ are computed in each band, and the autocorrelation functions from different frequency bands are combined. The two voicing measures differently express the magnitude of the combined autocorrelation.

This vector is computed at 5 ms intervals on all the speech data files. Figure 1 shows an example.

3.2. Building the Classifier

A Bayesian classifier is built on the differences between the feature vectors sampled 5 ms after the centers of the two sounds. The classifier assumes that the differences form a multivariate Gaussian distribution.

The classifier is trained to distinguish matched from unmatched sounds. By matched, we mean "in the same place in the same word", as determined in §2.1. These matched sounds will typically have the same phonological neighbourhood and often the same phonetic transcription. The training process finds the linear combinations of the feature vectors that are most effective at distinguishing the two classes. It suppresses linear combinations that are relevant to inter-speaker or utterance-to-utterance variability with the same text. The classifier thus makes those distinctions that are relevant to human perception of language.

The classifier assigns a difference to one or the other class on the basis of Equation 2:

$$\phi(\vec{a}, \vec{b}) = (\vec{a} - \vec{b}) \cdot M \cdot (\vec{a} - \vec{b})^T - \theta, \qquad (2)$$

$\vec{a}$ and $\vec{b}$ are audio description vectors, M is a matrix, and θ is a threshold that biases the classifier toward one class or the other. M and θ are chosen to minimise the total number of errors, which leads to nearly equal numbers of false negative and false positive errors. If $\phi < 0$, the two samples are most likely to have come from matched locations, and vice versa. This classifier is correct 69.6% of the time, with an 0.4% standard deviation, as measured from 28 random samples of a 25%/75% test-set/training-set split. (Chance performance would be 50%.)

3.3. Human Performance

We analyse human performance with a Bayesian Markov Chain Monte-Carlo analysis. This allows us to compute error bars for d' values in Signal Detection Theory (SDT) [20]. In this approach, we assume that the probability a subject will responds "S" is

$$P(S|d', b) = \Phi(d' \cdot c/2 + b), \qquad (3)$$

where d' measures the response to the stimulus and b is the group's overall bias toward responding "S". In Equation 3, Φ is the cumulative distribution function of a standard normal distribution and $c = -1$ or 1 depending on whether the particular sample is matched or unmatched. Bayes' Theorem then lets us reverse Equation 3 to compute a probability distribution for the parameters d' and b, given a set of observations. It is convenient to generate samples from this probability distribution by using the Metropolis algorithm [21]. We can then simply take the mean and standard deviation of d' for each sample to compute a confidence interval for d'.

Window width	Native speakers	L2 speakers	Machine
22 ms	1.06 ± 0.097	1.24 ± 0.09	0.92 ± 0.08
45 ms	0.96 ± 0.091	1.31 ± 0.10	1.19 ± 0.04
100 ms	1.60 ± 0.10	2.14 ± 0.11	1.34 ± 0.04

Table 1: D-prime values and standard errors for different subject groups at different window sizes.

4. Results and Discussion

Figure 2 summarises the results. Computing d' as described in §3.2 and §3.3 gives Table 1. Human performance increases dramatically between 45 ms and 100 ms windows in both groups ($P < 0.001$, $z \geq 3.8$), with a reduction in false positive rate by roughly half. No significant change is seen between 22 ms and 45 ms.

Machine performance is close to human performance for the two shorter windows, giving d' that is 80% of the average human value for a 22 ms window, and 104% as large for a 45 ms window. However, for the longer 100 ms window, human performance is substantially and significantly better than the classifier ($P < 0.025$ for native speakers, $P < 0.001$ for L2 speakers). This difference suggests that the human perceptual system is generating a richer representation of the sound in the longer windows, whereas the classifier isn't.

The L2 learners of English perform better at this task than native speakers. This is unexpected, as the sounds involved are British English. Unfortunately, the data are not strictly comparable, as the experimental conditions changed in mid-stream[2]. This difference in performance might result if the native speakers used a different definition of similarity than the L2 learners.

Figure 3 shows the effect of deleting individual components of the feature vector on the performance of the classifier. The classifier is built and tested with all components of the vector, then the process is repeated with the component under consideration missing. The change in performance is simply the average of the difference; more important components will make the change more negative. If the distribution of feature vector differences were multivariate normal, then the performance would always decrease; however the actual probability distribution of differences is longer-tailed and some projections of it are closer to square than ellipsoidal, so increases in performance upon deletion of a component are unsurprising. Individual spectrum components have a relatively small effect; that is because much of the information in each component can be found in its neighbours. The most important individual components are low frequency edges

[2] After the native speakers and one of the L2 speakers were studied, an equipment failure caused the computer, the room, and the location of the experimenter relative to the subject to be changed. However, none of the changes provide an obvious explanation for the difference in results.

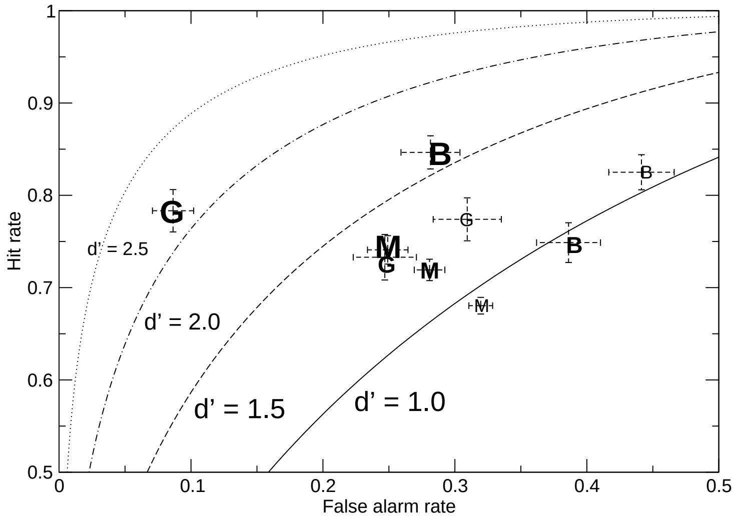

Figure 2: Human and machine performance on the same/different task. Points labelled "B" are from British subjects, "G" from Greek subjects, and "M" are results of machine classification. The horizontal axis is the false alarm rate, i.e. how often subjects judge different sounds to be the same, and the vertical axis is the hit rate, i.e. how often subjects judge equivalent sounds to be the same. The size of the letter shows the duration of the audio sample: smallest is for a 22 ms sample, medium for 45 ms, and large for 100 ms long samples. One standard-deviation error bars are shown. Contours of constant d' are plotted.

(e1) and voicing (V2).

5. Conclusion

On a task of deciding whether or not two sounds come from corresponding points in the same word, we have developed a classification algorithm whose performance is comparable to that of human subjects. The algorithm uses a feature vector based on a perceptual spectrum, but an edge detector on the 60–800 Hz band is very important, as is an autocorrelation-based measure of voicing.

6. References

[1] Baumgarte F., "Transfer function taken from "Improved Audio Coding Using a Psychoacoustic Model Based on a Cochlear Filter Bank", IEEE Transactions of Speech and Audio Processing, 10(7): 495–503, 2002.

[2] Plomp R. and Bouman M. A., "Relation between hearing threshold and duration for tone pulses", J. Acoustical Society of America, 31(6):749–758, 1959.

[3] Shen J.-L. and Hung S. H. and Lee L. -S., "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments", http://www.ee.columbia.edu/ dpwe/papers/ShenHL98-endpoint, International Conference on Spoken Language Processing, 1998.

[4] Meddis R. and O'Mard L.,"A unitary model of pitch perception", J. Acoustical Society of America, 102(3):1811–1820, 1997.

[6] Linkai Bu and Tzi-Dar Chiueh,"Perceptual Speech Processing and Phonetic Feature Mapping for Robust Vowel Recognition", IEEE Transactions on Speech and Audio Processing, 8(2):105–114, 2000.

[8] Flanagan J. L. and Saslow M. G. "Pitch Discrimination for Synthetic Vowels", J. Acoustical Society of America, 30(5):435–442, 1958.

[9] Klatt D. H. "Discrimination of Fundamental Frequency Contours in Synthetic Speech: Implications for Models of Pitch Perception", J. Acoustical Society of America, 53(1):8-16, 1973.

[10] Stoll G. "Spectral-pitch pattern: A concept representing the tonal features of sounds", in M. Clynes (ed), "Music, Mind and Brain", pp. 271–278, Plenum Press, New York, 1982.

[11] Houstma A. J. M. and Rossing T. D., "Effects of signal envelope on the pitch of short complex tones",

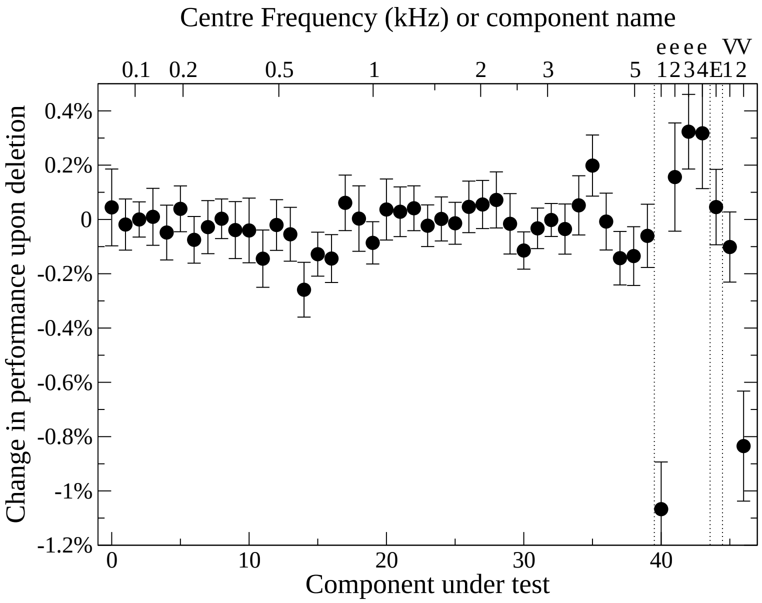

Figure 3: Effect of deleting individual components of the feature vector. The vertical axis is the change in performance of the classifier, averaged over the 28 training/test splits of the data (error bars show the standard deviation of the test set results). The horizontal axis shows the 47 components of the feature vector: 40 spectrum components, 4 edge detector components, 1 spectral entropy measure and the 2 voicing components, with dotted lines as separators.

J. Acoustical Society of America, 81(2):439–444, 1986.

[13] Pols L. C. W. and Schouten M. E. H. “Perception of tone, band and formant sweeps”, in Schouten M. E. H. (ed) The Psychophysics of Speech Perception, NATO ASI Series, 1987.

[14] Howell P. and Darwin C. J. “Some properties of auditory memory for rapid formant transitions”, Memory and Cognition, 5(6):700–708, 1977.

[15] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective Measures of Speech Quality*. Prentice-Hall, Englewood Cliffs, New Jersey, 1988.

[16] Fletcher H. and Munson W. A., “Loudness, its definition, measurement, and calculation”, J. Acoustical Society of America, 5:82–108, 1933.

[17] S. Wang, A. Sekey, and A. Gersho. An objective measure for predicting subjective quality of speech coders.

[18] Stevens S. S., “Perceived Level of Noise by Mark VII and Decibels”, J. Acoustical Society of America, 51(2-2):575–602, 1971.

[19] Slater A. and Coleman J. “Non-segmental analysis and synthesis based on a speech database”, in Bunnel H. T. and Idsardi W. (eds), Proceedings of ICSLP 96, Fourth International Conference on Spoken Language Processing, 4:2379–2382, 1996.

[20] Macmillan N. A. and Creelman C. D. *Detection Theory: A User’s Guide* Laurence Erlbaum Associates, London, second edition 2005. ISBN 0-8058-4231-4.

[21] “Practical Markov Chain Monte Carlo” Statistical Science, pp. 473–483, 1992.