# Testing the Ecological Validity of Repetitive Speech

*Greg Kochanski and Christina Orphanidou*

The University of Oxford
greg.kochanski@phon.ox.ac.uk and christina.orphanidou@linacre.oxford.ac.uk

## ABSTRACT

Can one trust experiments conducted with repetitive speech to represent normal language behaviour? We compare the spectra of repetitive productions of phrases with the same phrases read from a randomised list. We use a data-driven spectral distortion measure that is trained to respond to linguistically relevant differences: it is based on a classifier that separates sounds into linguistically equivalent or not. We find that repetitive speech is not distinct from individually uttered speech. The difference between these two sorts of speech is smaller than variation within each. It is substantially smaller than typical differences between utterances produced by different subjects.

**Keywords:** production phonetic style machine learning

## 1. INTRODUCTION

Repetitive speech[1] has been used in a number of experiments on speech rhythm and syllable structure (e.g. [12, 4]). It is also necessary for stroboscopic MRI imaging of the vocal tract [9, 14] and other experiments. Such research assumes that repetitive speech – which is uncommon in human interactions – is representative of natural speech activities. We investigate this by constructing a linguistically oriented signal-processing measure of the difference between utterances.

A psychophysical distance measure exists for loudness [6, 17], but a more general distance measure that distinguishes changes in vowel quality or timbre is desirable. Such algorithms have been developed to evaluate speech coders ([18, 13] and references therein). We extend this work with a data-driven approach, where the distance measurement can be tuned to respond to linguistically relevant differences. We use this measure to assess the difference between repetitive and non-repetitive readings of the same text, compared to intra-speaker, inter-speaker and inter-text variation.

## 2. Experimental Methods

We used two sets of data, one for training the distance measurements and another, the science data, containing the contrasts to be investigated. All participants were linguistically naive speakers of Standard Southern British English.

The data for training the distance measurement were a set of phonetically rich sentences that ten subjects, aged 19–62, read from randomised lists. A total of 2578 utterances were produced from 828 different texts. The mean sentence length was 6.5 words. Some of the recordings were hand-segmented into sentences and some was automatically recorded as individual sentences. Utterance beginnings and endings were marked automatically using HTK [3], with manual checking and cleaning.

The science data were collected from nine subjects age 19–33 (no subjects in common with the training set). This experiment involved reading a set of 48 phrases, 4–6 syllables long. phrases were selected for broad coverage of phonemes, minimal duplication of words, and no unusual vocabulary.[2]

The science data were collected under two conditions. For the first condition ("list"), the subjects read a randomised list, with 5 copies of each of the 48 phrases, and 24 filler phrases.

After this, a training task gave the subjects practice on reading these phrases repetitively. For each subject, two blocks of 12 phrases were randomly selected. Subjects were asked to read each phrases ten times in a row. One block was simply read out, then the second block was read while the subjects tapped their finger to the prominent syllables.

In the second experimental condition ("repeat"), the subjects read each of the remaining 24 phrases, repeatedly, to the beat of a metronome. (Subjects had earlier picked two comfortable rates; half were read at each rate. Rates were $84 \pm 8$ beats per minute, and the two rates typically differed by 4 bpm.

The audio recordings were inspected for clipping and spurious noises (e.g. coughs, rustling paper), and approximately 5% were rejected for those reasons. A further 1% were rejected for gross mispronunciations.

## 3. Analysis

The analysis consists of two stages: first, constructing the distance measurement using the training data, and second, the analysis of the science data.

### 3.1. Constructing the Distance Measurement

The training process begins by computing an acoustic description vector. Most of the vector is computed from a "perceptual spectrum" [2].[3] This representation, $R(t, \omega)$, is computed by feeding the speech signal into a filter bank spanning 60–6000 Hz. Filters are taken from [1] and spaced every 0.7 erb. The filtered outputs are half-wave rectified, then a cube root is taken to approximate the perceptual loudness response.

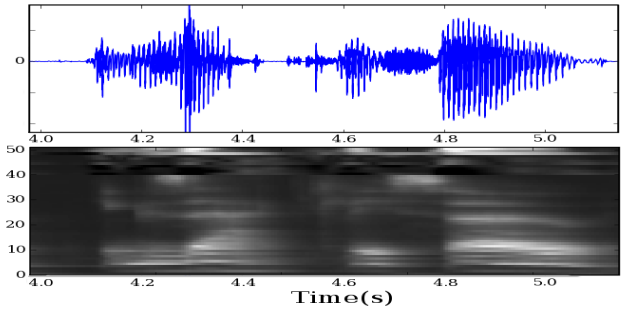


**Figure 1:** The top panel shows sample audio signal for one utterance, "Exactly so." The bottom panel shows the audio description vectors as a grey scale. Vectors run vertically with the spectrum components at the bottom and voicing estimates on top.

The acoustic description vector has 51 components:

- 1-40 A normalized spectrum, obtained by smoothing $R(t, \omega)$ in the time domain with a single-pole low-pass filter (time constant from [11]). This covers 300–3000 Hz. It is then normalized by a spatial and temporal average of $R(t, \omega)$.
- 41-48 A part designed to detect changes in the spectrum. The above normalized spectrum is collected into four broad frequency bins and then differentiated over a time scale of 40 ms. An additional set of four components is obtained by taking the absolute values.
- 49 One component is a spectral entropy measure, inspired by [15], and computed from $R(t, \omega)$.
- 50, 51 Two final components are indicators of voicing, inspired by [10]. Autocorrelations of $R(t, \omega)$ are computed in each frequency band, and then combined. The two voicing measures combine the autocorrelations differently.

This vector is computed at 5 ms intervals on all the speech data files. Figure 1 shows an example. Based on informal tests, our results are not strongly dependent on the details of the acoustic description vector.

Next, we construct a classifier from which we shall extract the distance measurement. The classifier is trained to decide whether a pair of points in two utterances are lexically equivalent or not.

Feature vectors used to train the classifier are obtained by randomly pairing utterances (93% of the time from different speakers). We match the utterances using the acoustic description vectors as input to the dynamic time warping (DTW) algorithm from [16]. If the utterances come from the same text, matching points will typically be in the same place in the same word. These lexically equivalent locations will typically have the same phonological neighbourhood and often the same phonetic transcription.

Next, we subtract the acoustic description vectors at matching points. These differences form the "equivalent" class. "Nonequivalent" feature vectors are constructed by subtracting acoustic descriptions that are at least 250 ms away from alignment. From 37,000 pairs of utterances, we sample 80,000 pairs of locations for each of the two classes.

The training process finds the linear combinations of the feature vectors that are most effective at distinguishing the two classes. It will suppress linear combinations that have much inter-speaker or intra-speaker variation. The classifier thus makes distinctions that are relevant to human perception of language. The resulting distance measurement shows how different the sounds are, emphasising lexically important differences.

The classifier is Bayesian and assigns one or the other class on the basis of Equation 1:

$$(1) \quad \phi(\vec{a}, \vec{b}) = (\vec{a} - \vec{b}) \cdot M \cdot (\vec{a} - \vec{b})^T - \theta,$$

where $\vec{a}$ and $\vec{b}$ are audio description vectors, $M$ is a matrix, and $\theta$ is a threshold that biases the classifier toward one class or the other. $M$ and $\theta$ are set, based on the training data. This follows [8, 5], simplified because the centres of the two classes are both at the origin. If $\phi < 0$, the two samples are most likely to have come from lexically equivalent locations, and vice versa.

One expects that the acoustic differences between lexically nonequivalent locations are larger than differences between equivalent locations. Thus there should not be many large (in the absolute sense), negative eigenvalues of $M$. This is indeed the case: the largest negative eigenvalue is the $12^{\text{th}}$ largest and is only 16% as big as the largest positive one.

This classifier is correct $83 \pm 0.4\%$ of the time for a test set (using the Quadratic Forest code from [7, §II.H]), where chance performance would be 50%.

If all the eigenvalues of a quadratic form such as Equation 1 were positive,

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

would be a distance measure. It would be zero for two identical feature vectors, positive otherwise, and obey the triangle inequality. We can arrange for this to be the case, by deleting the eigenvalues that are negative or smaller than 6.5% of the largest eigenvalue. (Here, we follow [20].) This leads to a lower-rank approximation to $M$ and a classifier operating in a 19-dimensional (vs. 51) space. This simpler classifier has an accuracy of 80.7%, showing that we have not lost much information.

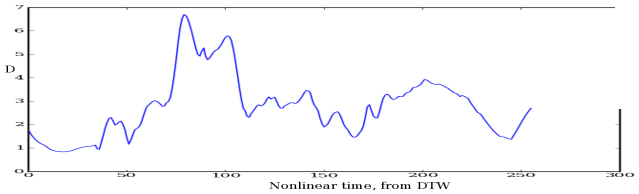


**Figure 2:** Distance vs. a nonlinear time axis (the result of the dynamic time warping algorithm) for one pair of utterances.

We use this simplified, positive-definite $M$, to quantify the distance between utterances. Note that there is a $\vec{a}$ and $\vec{b}$ to describe the acoustic properties at each point in an utterance, and thus $D(\vec{a}, \vec{b})$ is a function of time. A sample curve of distance vs. time is shown in Figure 2.

### 3.2. Computing Distances

For the science data, we have five sets of pairs of utterances to be compared:

- **sCsSsT**: Pairs that have the **s**ame experimental **C**ondition, the **s**ame-**S**peaker and **s**ame-**T**ext.
- **dCsSsT**: Pairs of utterances from **d**ifferent **C**onditions, the **s**ame **S**peaker and the **s**ame **T**ext.
- **sCdSsT**: Pairs from **d**ifferent **S**peakers, but which are otherwise the **s**ame.
- **sCsSdT-dtw**: Same condition and speaker, but **d**ifferent **T**exts, using the DTW alignment.
- **sCsSdT-linear**: As above, but rather than using the DTW alignment, we use a robust linear fit to the DTW alignment.

Figure 3 displays histograms of the distances between utterances defined by Equation 2 in each comparison set. Each element of the histogram is a time-average of the distance between a pair of utterances.

## 4. Results and Discussion

We can test the distance measure by checking that it agrees with reasonable expectations. Pairs of different texts (lowest panels) should give larger distances than identical texts (**sCsSsT**, top). Similarly for speakers: $D(\textbf{sCdSsT}) > D(\textbf{sCsSsT})$. Further, text differences should be more important than speaker

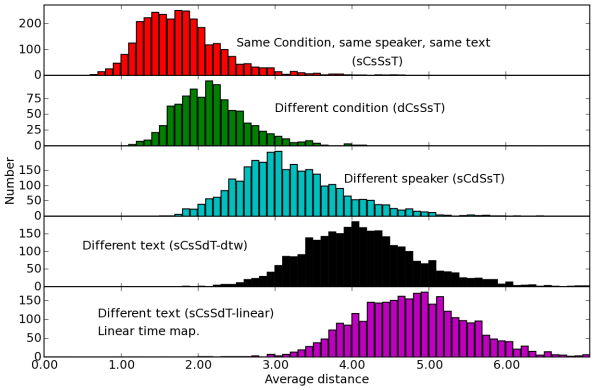


**Figure 3:** Histograms of distances. Each measurement is the average distance of a pair of time-aligned utterances. Each sub-figure corresponds to a different relationship between the utterances in the pair.

differences: $D(\textbf{sCsSdT-dtw}) > D(\textbf{sCdSsT})$, and $D(\textbf{sCsSdT-linear}) > D(\textbf{sCdSsT})$. (Here $D(X)$ refers to the distances between pairs of utterances in set $X$.) Also, utterances under different experimental conditions (**dCsSsT**) should be further apart than utterances under the same experimental condition (**sCsSsT**). All these expectations are supported and the means of the distributions are significantly different at the $P << 10^{-6}$ level via a $t$-test ($t > 20$, $> 1100$ degrees of freedom). They are even fairly well supported on a single-sample basis: the probabilities that these statements are true for randomly chosen pairs of pairs ranges from 74% up to 99.8%.

The distribution of distances in Figure 3 is fairly narrow. In the **sCsSdT** case, the standard deviation is only 15% of the mean, and the standard deviation among the **sCsSsT** cases is only 11% of the mean distance between utterances with different texts. This suggests that the accuracy of our distance measurements is good. In fact, given that our corpus is short phrases, with a variety of metrical and phonological patterns, the variance seems surprisingly small.

### 4.1. Comparison between repetitive and list speech

The **dCsSsT** distance histogram in Figure 3 (second panel) compares repetitive speech to speech from randomised lists. The mean **dCsSsT** distance is smaller than the mean inter-speaker distance (**sCdSsT**), significant at better than the $P < 10^{-6}$ level ($t = 18$, 1100 degrees of freedom). In fact, 90% of all $D(\textbf{sCdSsT})$ measurements are larger than a randomly chosen $D(\textbf{dCsSsT})$ measurement.

To estimate the size of the effect, we note that our distance measure (Equation 2) is Euclidean and

mathematically equivalent to a variance. Therefore, if **dCsSsT** contains the same utterance-to-utterance variability as **sCsSsT**, we can compute how much changing the experimental condition changes the acoustic properties of the utterance via Equation 3:

$$\Delta C = \left(D^2(\mathbf{dCsSsT}) - D^2(\mathbf{sCsSsT})\right)^{1/2}. \quad (3)$$

Similarly, we can compute the effect of changing speaker or text.

We find that $\Delta C = 1.3$, on the scale of Figure 3, so that a change in experimental condition has an effect which is 71% of the utterance-to-utterance variability, or 40% as large as a change of speaker, or 26% as large as a change of the text.

## 5. Conclusion

We have developed a data-driven technique for measuring a linguistically-relevant distance between utterances. It behaves properly in five test cases. This technique shows promise for quantitative studies of dialect differences and speech synthesis. In terms of linguistically relevant changes in the spectrum, repetitive speech does not seem much different from that produced by the standard laboratory technique of asking subjects to read lists of randomised phrases. The difference between repetitive speech and random-list speech is smaller than utterance-to-utterance variability and substantially smaller than typical differences between speakers.

We gratefully acknowledge the the UK's Economic and Research Council for funding via RES-000-23-1094.

## 6. REFERENCES

[1] F. Baumgarte. Improved audio coding using a psychoacoustic model based on a cochlear filter bank. *IEEE Transactions on Speech and Audio Processing*, 10(7), October 2002.

[2] L. Bu and T.-D. Chiueh. Perceptual speech processing and phonetic feature mapping for robust vowel recognition. *IEEE Transactions on Speech and Audio Processing*, 8(2):105–114, 2000.

[3] Cambridge University Engineering Department. *HTK Speech Recognition Toolkit*, January 2007.

[4] K. J. de Jong, B.-J. Lim, and K. Nagao. The perception of syllable affiliation of singleton sptops in repetitive speech. *Language and Speech*, 47(3):241–266, 2004.

[5] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2000.

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

[7] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner. Loudness predicts prominence: Fundamental frequency lends little. *J. Acoust. Soc. of America*, 118(2):1038–1054, 2005.

[8] D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4(5):720–736, 1992.

[9] K. Mathiak, U. Klose, H. Ackermann, I. Hertrich, W. E. Kineses, and E. Grodd. Stroboscopic articulography using fast magnetic resonance imaging. In *Proceedings of the Fifth Seminar on Speech Production: Models and Data*, pages 97–100, 2000. Munich: Universität München.

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

[11] R. Plomp and M. A. Bouman. Relation between hearing threshold and duration for tone pulses. *J. Acoustical Society of America*, 31(6), June 1959.

[12] R. F. Port. Meter and speech. *J. Phonetics*, 31:599–611, 2003.

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

[14] C. H. Shadle, M. Mohammad, J. N. Carter, and P. J. B. Jackson. Dynamic magnetic resonance imaging: new tools for speech research. In *Proceedings of 14th Int. Cong. Phon. Sci.*, pages 623–626, 1999.

[15] J.-L. Shen, J.-W. Hung, and L.-S. Lee. Robust entropy-based endpoint detection for speech recognition in noisy environments. In *International Conference on Spoken Language Processing*, 1998.

[16] A. Slater and J. Coleman. Non-segmental analysis and synthesis based on a speech database. In H. T. Bunnell and W. Idsardi, editors, *Proceedings of ICSLP 96, Fourth International Conference on Spoken Language Processing*, volume 4, pages 2379–2382, 1996.

[17] S. S. Stevens. Perceived level of noise by Mark VII and decibels. *J. Acoustical Society of America*, 51(2, part 2):575–602, 1971.

[18] S. Wang, A. Sekey, and A. Gersho. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*, 10(5):819–829, June 1992.

[19] E. D. Young and M. B. Sachs. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers. *J. Acoustical Society of America*, 66(5):1381–1403, 1979.

[20] S. A. Zahorian and M. Rothenberg. Principal components analysis for low-redundancy encoding of speech spectra. *J. Acoustical Society of America*, 69(3):832–845, March 1981.

[1] E.g. a subject rhythmically reading "Under the desk. Under the desk. Under the desk…" to a metronome.

[2] Representative phrases in the science data are "Nothing matters," "Talking of wandering," and "Not to my knowledge."

[3] This is a phenomenological model; actual signals in the auditory nerve are not so simple [19]. We deviate from [2] in that we neglect masking.