Published in Proceedings of the International Conference on Spoken Language Processing, Jeju, Korea, Oct 2004.

ACCDIST: a Metric for Comparing Speakers' Accents

Mark Huckvale

Department of Phonetics and Linguistics
University College London, U.K.
m.huckvale@ucl.ac.uk

Abstract

This paper introduces a new metric for the quantitative assessment of the similarity of speakers' accents. The ACCDIST metric is based on the correlation of inter-segment distance tables across speakers or groups. Basing the metric on segment similarity within a speaker ensures that it is sensitive to the speaker's pronunciation system rather than to his or her voice characteristics. The metric is shown to have an error rate of only 11% on the accent classification of speakers into 14 English regional accents of the British Isles, half the error rate of a metric based on spectral information directly. The metric may also be useful for cluster analysis of accent groups.

1. Introduction

A speaker's accent marks him or her as a member of a group. These groups have been defined by geographical areas, by socio-economic class, by ethnicity, or for second language speakers, by the identity of the speaker's first language. As listeners sensitive to accents we can tell whether a speaker belongs to our group, and as talkers we adapt our speech when we want to belong (or appear to belong) to a different group [1]. For this to be possible it must be the case that accents are stylised patterns of speaking that recur across members of the group. We assume these patterns affect word frequency, the phonological coding used in the pronunciation lexicon, the phonetic realisation of phonological units and the prosody of utterances [2].

Speech technology has yet to deal adequately with pronunciation variation across accent groups. In speech recognition, a mismatch in accent between the speakers used in testing and training can lead to a 30% increase in word error rate [3]. In speech synthesis, synthetic voices are fixed in one accent due to the increasing use of corpus-based synthesis methods [4] operating from the speech of a single speaker.

There are a number of reasons that make accent variation difficult for contemporary speech modelling techniques:

1. Accent variation is not just a shift in phonetic realisation: accents differ in their inventory of phonological segments and their distribution in the lexicon. This means that keeping one dictionary and adapting the mean spectral realisations of phone models is insufficient.
2. Phonetic variation can involve large spectro-temporal changes in realisation: for example, monophthongs can become diphthongs, plosives can become fricatives, and segments can be inserted and deleted. Phone models which are good models of spectro-temporal variation of a phonological unit in one accent may be poor models in another. This means that adapting the dictionary but keeping one set of phone models is also not sufficient.
3. Databases of speech used for training recognisers are not well controlled for accent: it is likely that any given phone model is trained with speech from a number of accent groups. Such impure models confuse attempts at dealing with accents by phonetic and phonological adaptation. A model of /ɑː/ containing both [æ] and [ɑː] may be useful for modelling "bath" but not "palm".
4. Sociolinguists define accent groups according to convenient cultural indicators rather than on the basis of similarity: it is unlikely that all the known groups are necessary or sufficient. In addition, because the groups are not defined by objective similarity, it is hard to find a representative sample of speakers of an accent.
5. Accent variation is only one component of variability of a speaker: speakers also differ according to their age, size, sex, voice quality, speaking style or emotion, and recordings are affected by environment, background noise and the communication channel. But since accent is a characteristic of a *group* of speakers, it is hard to control these other influences.

Thus speech technology could benefit from modelling techniques which are sensitive to the particular character of accent variation. Better modelling of accents would allow recognition systems to accommodate speakers from a wide range of accents, including second language speakers. A better understanding of the acoustic-phonetic structure of accents might lead to means for morphing voices across accents [5] which could allow concatenative synthesis systems to speak in multiple accents. Finally, better definitions of accent groups could lead to new sociolinguistic insights into how groups form and change.

2. Approaches to Modelling an Accent Group

2.1. Global acoustic distribution

The simplest way to characterise an accent group is to make a model of the probability distribution of the acoustic vectors recorded from a set of speakers from one group. For example, Huang et al [3] modelled four regional accents of Mandarin using a Gaussian mixture model with 32 components to model the pdf of spectral envelope features from 1440 speakers. Accent recognition can then be performed without using a known text or requiring phonetic labelling: Huang et al achieved an accent recognition rate of 85% using gender-dependent models. However, such a global model seems to be a crude way to model differences in phonetics and phonology, particularly when the models also contain other speaker variability.

2.2. Accent-specific phone models

Having known text read by speakers of known accent groups allows the building of a set of phone models for each accent. The models can be used in accent recognition simply by finding which phone set gives the highest probability to any unknown test utterance. For example, Teixeira et al [6] obtained about 65% accent recognition rate for five foreign accented English speaker groups. The weakness of this approach is that phonological variation is not exploited, since the recognisers do not necessarily use the same best phone transcription for the utterance. When the text and a phonological transcription is known, the accent can be found using the same phone sequence for all sets and performance is much higher. For example, Arslan & Hansen [7] obtained a 93% accent recognition rate for four foreign accented English speaker groups. However, such an approach requires that sufficient data be available to build phone models, and that this data come from a range of speakers so as to accommodate speaker variability. Thus it assumes that accent groups are known and that training speakers can be assigned to groups.

2.3. Analysis of pronunciation system

While accent recognition based on accent-specific phone models works well for a small number of varieties of foreign-accented English, it is not clear that the technique would scale well to the problem of dealing with a larger number of more similar regional accents of a language. We believe a more sensitive technique could come from a study of a speaker's pronunciation system rather than his acoustic quality. Barry et al [8] developed a regional accent recognition technique based on acoustic comparisons made *within* one known sentence. Formant frequency differences between vowels in known words were used to assign the speaker to one of four English regional accents with an accuracy of 74%.

Barry's idea to look at the relationship between the realisations of known segments rather than their absolute spectral quality was recently advanced further by the work of Nobuaki Minematsu [9]. His idea was to perform cluster analysis on a set of phone models for a single speaker, then study the resulting phonetic tree to establish the pronunciation habits of the speaker. By this, Minematsu hoped to identify where the speaker's pronunciation differed to some norm. However in this paper we take Minematsu's idea a step further, and apply it to the problem of accent characterisation and recognition. We use the similarities between segments to characterise the pronunciation system for a speaker, then compare his pronunciation system with average pronunciation systems for known accent groups to recognise his accent. We first describe the experimental data and baseline results.

3. Accent Data and Baseline Performance

3.1. Speech data

Speech material was extracted from the Accents of the British Isles (ABI) corpus [10]. Ten male and ten female speakers from 14 accent areas (see Table 1) spoke the same 20 short sentences. Six speakers who did not complete enough of the set were excluded, leaving 274 speakers.

Table 1: ABI Corpus Accent groups and codes

Code	Accent	Code	Accent
brm	Birmingham	lvp	Liverpool
crn	Cornwall	ncl	Newcastle
ean	East Anglia	nwa	North Wales
eyk	East Yorkshire	roi	Dublin
gla	Glasgow	shl	Scottish Highlands
ilo	Inner London	sse	South East
lan	Lancashire	uls	Ulster

A phonological transcription was generated for each sentence using Southern British English pronunciations, and phonetic segmentation was performed using forced alignment with the HTK Hidden Markov Modelling toolkit [11]. All subsequent analysis was made using only the vowel segments in the 20 sentences including diphthongs but excluding schwa. This gave between 100 and 140 vowels per speaker (some speakers did not complete some sentences).

3.2. Formant frequency distance metric

Baseline accent recognition performance was first obtained using a metric based on formant frequencies for the vowels. Formant frequency estimation was performed using the *formanal* program of SFS [12]. Each vowel was divided into two halves by time, and the median value of the first four formant frequencies in each half were combined to build an 8-dimensional vector for classification.

Accent recognition performance was estimated by taking the mean formant frequency vector for each vowel in the 20 sentences for each group excluding the speaker under test, then determining the closest group from the mean euclidean distance between the test speaker's vowels and the accent group mean vowels. This process did not require phonological labelling of the vowels since only vowels occurring in the same words were matched with each other. This procedure was then repeated over each speaker in turn. Accent recognition accuracy is given in Table 2. This table also shows the effect of the gender of the speakers. Performance was measured using the means of all speakers, of speakers of the same sex as the test speaker, and of speakers of the opposite sex to the test speaker. Unsurprisingly, performance is slightly improved when same-sex models are used, and significantly worsened when other-sex models are used. This shows that the metric is significantly sensitive to speaker characteristics unrelated to accent.

Table 2: Formant metric performance

Speaker set	Accent recognition rate
Any sex	51.1%
Same sex	59.1%
Other sex	35.4%

This result can be improved by standardising the formant frequency values to a unit normal distribution using the mean and variance for each speaker independently. Recognition was then performed as before, using leave-one-out, a mean euclidean distance to the accent group mean vowels, matching vowels by word context, and for three gender conditions. The recognition results are shown in Table 3. Although there is a significant increase in accuracy over the un-normalised condition, and a much smaller effect of the same-sex models

as expected, there is still a significant drop in performance in the other-sex condition, showing that formant frequency normalisation alone does not compensate for gender differences.

Table 3: Normalised formant metric performance

Speaker set	Accent Recognition Rate
Any sex	71.9%
Same sex	72.6%
Other sex	59.1%

3.3. Spectral envelope distance metric

To obtain a baseline performance for a spectral envelope metric, mean spectral envelopes for each vowel were obtained as follows. Each sentence was analysed using a 19-channel auditory filterbank designed using the specification of Holmes [13]. The mean of each frame was subtracted and added as a 20^{th} value. Each vowel was divided into two halves by time, and the mean spectral envelopes in each half were combined to form a 40-dimensional vector for classification.

Accent recognition performance was then measured as before, using leave-one-out, a mean euclidean distance to the accent group mean vowels, matching vowels by word context, and for three gender conditions. Accent recognition accuracy is given in Table 4. Performance in the "any sex" and "other sex" conditions are similar to the normalised formant frequency metric although the spectral metric has no frequency normalisation. Better performance is obtained in the "same sex" condition. Again the results show a sensitivity to absolute speaker characteristics not just accent.

Table 4: Spectral metric performance

Speaker set	Accent recognition rate
Any sex	71.5%
Same sex	79.2%
Other sex	54.7%

4. ACCDIST Metric

4.1. Metric description

The baseline results confirm that a metric based on absolute spectral properties of the speech is affected by characteristics of the speaker other than their accent. This leads us to conclude that a metric based on relative measures made within a speaker could provide better performance

Accent Characterisation by Comparison of Distances in the Inter-segment Similarity Table (ACCDIST) is a metric based on the form of a speaker's accent as expressed in the relative similarity of his or her segment realisations with each other. As an example, consider the two distance tables for the stressed vowels in "after", "father" and "cat" spoken by a Birmingham speaker and a South-east British speaker shown in Table 5.

These distance tables reflect the fact that the vowel in “after” for the Birmingham speaker was more similar to his vowel in “cat”, while for the South-east speaker it was more like his vowel in “father”. Note that a comparison of these two tables highlights a difference in pronunciation system without requiring us to compare absolute spectral qualities across speakers.

Table 5: Example vowel distance tables

Birmingham		
Distance	Father	cat
after	3.48	2.14
father	0.00	3.62

South-east		
Distance	father	cat
after	2.27	3.21
father	0.00	3.71

The key to the ACCDIST metric then, is the calculation of the *correlation* between a pair of such segment distance tables. In practice, the tables are much larger and could include all vowels and consonants for a speaker, although in this paper we have used only vowels. A correlation measure is chosen as this makes the comparison insensitive to the absolute distances between segments for a speaker which may also vary with the speaking style and voice quality.

Importantly the ACCDIST metric can be made sensitive to both phonetic and phonological changes in an accent: we simply calculate the distance tables from single segment realisations, use the same text for both speakers, and only match vowels across speakers when they occur in the same words. This way we label vowels as “the vowel in ‘cat’” rather than as /æ/. The only assumptions made are that the two speakers have spoken the same words and that the words contain corresponding sub-components.

4.2. Recognition to accent mean

We first evaluate the ACCDIST metric using a similar procedure as before. Accent recognition performance was estimated by taking the mean of the distance tables calculated across all 140 vowels for each speaker of each accent group excluding the speaker under test, then determining the closest group from the correlation between the test speaker vowel distance table and the accent group mean tables. As before, only vowels occurring in the same words were matched with each other. This was then repeated over all speakers in turn. Accent recognition accuracy is given in Table 6. What is interesting here is that not only is accent recognition performance considerably higher than the baseline, but that there is much less dependency on gender, with even the cross-gender recognition rate being over 80%.

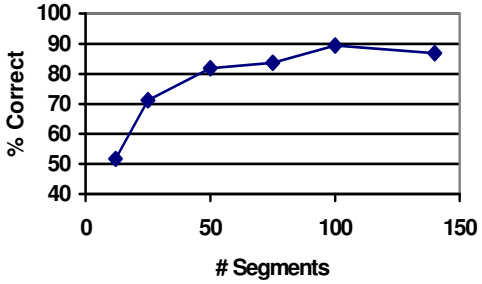
Table 6: ACCDIST metric performance

Speaker set	Accent Recognition Rate
Any sex	86.9%
Same sex	87.2%
Other sex	81.4%

To explore how many vowels are required for good accent recognition performance, random subsets of the distance tables with 100, 75, 50, 25 and 12 vowels were evaluated. Recognition rate as a function of distance table size is shown in Figure 1. Interestingly, the best performance of 89.4% occurs at 100 vowels, but performance is still better than 80% with only 50 vowels. It is possible that better performance can

be obtained for fewer vowels by careful selection of vowel type.

Figure 1: ACCDIST metric performance with # segments



4.3. Pairwise comparisons

For some applications, it would be useful to compare a speaker not to a group mean but to another speaker. To evaluate how well the ACCDIST metric performs without averaging the distance tables, accent recognition rate was calculated using the individual distance tables and a one-nearest neighbour decision rule. The recognition results are shown in Table 7. Here performance is still very high, greater than 80% correct, and dependency on gender is still small.

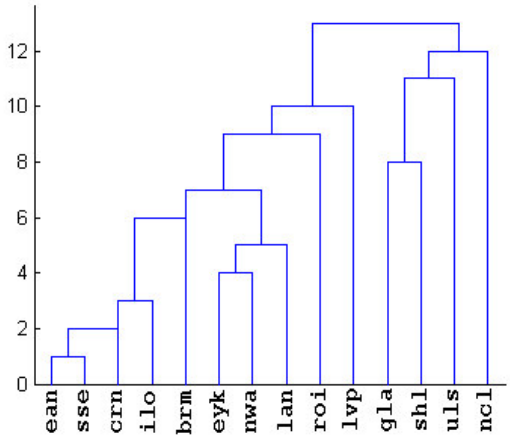
Table 7: ACCDIST metric performance using 1-nearest neighbour

Speaker set	Accent Recognition Rate
Any sex	80.3%
Same sex	80.7%
Other sex	75.9%

4.4. Accent mean clustering

ACCDIST may be particularly useful in cluster analysis of accents, since this will mean that we no longer need to rely on accent labels used by sociolinguists. To demonstrate cluster analysis, the fourteen accent groups were clustered into a dendrogram using the ACCDIST metric, and the results shown in Figure 2.

Figure 2: Clustering of mean accent groups



What is interesting in the cluster analysis is that the Southern British Isles (ean, sse, crn, ilo), Northern British Isles (eyk, nwa, lan) and Scottish (gla, shl) regional accents cluster separately, just as one might predict. There are also intriguing similarities between Ulster and Scottish accents, and Liverpool and Dublin accents: groups which have real historical connections. A similar patterning was observed when individual speakers rather than groups were clustered.

5. Conclusions

This paper introduced a new metric for the comparison of accents. The metric is based on the correlation of segment distance tables which avoids the problem of comparing absolute spectral characteristics across speakers. The metric has been shown to perform well on a difficult regional accent recognition task, showing better performance and less sensitivity to gender than direct spectral measures.

6. Acknowledgements

I would like to thank Nobuaki Minematsu for introducing me to the idea of speaker-dependent phone clustering; and Paul Iverson for suggesting correlation as a suitable similarity measure. Thanks to 2020Speech Ltd for making the ABI corpus available for research purposes.

7. References

1. Giles, H. & Powesland, P.F., *Speech Style and Social Evaluation*, Academic Press, London, 1975.
2. Wells, J.C., *Accents of English*, Cambridge University Press, 1982.
3. Huang, C., Chang, E. & Chen, T., "Accent Issues in Large Vocabulary Continuous Speech Recognition", *Microsoft Research China Technical Report*, MSR-TR-2001-69, 2001.
4. Taylor, P.A., and Black, A.W., "Concept-to-speech synthesis by phonological structure matching", *Proc. EuroSpeech-99*, 623-626, 1999.
5. Ho, C-H., Vaseghi, S. & Chen, A., "Voice conversion between UK and US accented English", *Proc. EuroSpeech-99*, 2079-2082, 1999.
6. Teixeira, C., Trancoso, I. & Serralheiro, A., "Accent identification", in *Proc. ICSLP'96*, 1784-1787, 1996.
7. Arslan, L.M., & Hansen, J.H.L., "Language accent classification in American English", *Speech Communication*, 18, 353-367, 1996.
8. Barry, W.J., Heoquist, C.E. & Nolan, F.J., "An approach to the problem of regional accent in automatic speech recognition", *Computer Speech and Language*, 3, 355-366, 1989.
9. Minematsu, N. & Nakagawa, S., "Visualization of Pronunciation Habits Based upon Abstract Representation of Acoustic Observations", *Proc. Integration of Speech Technology into Learning 2000*, pp.130-137, 2000.
10. http://www.aurix.com/
11. http://htk.eng.cam.ac.uk/
12. http://www.phon.ucl.ac.uk/resource/sfs/
13. Holmes, J.N., "The JSRU channel vocoder", Proc. IEE, 127, Pt. F, 53-60, 1980.