Huckvale, M., "The New Accent Technologies: Recognition, Measurement and Manipulation of Accented Speech", in Research and Application of Digitized Chinese Teaching and Learning, ed. By P. Zhang, T.-W. Xie, S. Lin, J.-H. Xie, A.C. Fang, and J. Xu. Beijing: Language and Culture Press. pp 28-37, 2006.

The new accent technologies: recognition, measurement and manipulation of accented speech

Mark Huckvale
Phonetics and Linguistics
University College London
M.Huckvale@ucl.ac.uk

Abstract

Advances in speech technology, speech signal processing and phonetic representation are leading to new applications within Accent studies. These technologies will allow us to automatically identify the features of an accent, to cluster speakers into accent groups, to adapt our pronunciation dictionaries on-line to a speaker's accent, to measure the similarity between accents, even to modify recordings of a speaker to change their accent. These technologies apply to both regional and foreign accented speech and have considerable potential in language learning. For example they will allow a learner's accent to be evaluated and diagnosed, they will allow the demonstration of pronunciation targets in the learner's voice, and they can improve the intelligibility of foreign accented speech to native listeners.

In this article I will describe some of the underlying components of the new accent technologies and demonstrate their use. In speech recognition, I will show how an accent feature system can be used for pronunciation dictionary adaptation to improve recognition performance without the need to identify the accent of the speaker. In experimental phonetics, I will show how measures of self-similarity provide a means to measure and evaluate accent independently of speaker characteristics. In speech signal processing, I will show how accent morphing techniques can be used to modify a speaker's accent in a given recording, and show how such methods can lead to an increase in the intelligibility of foreign accented speech to native listeners.

1. Introduction

Speech technology has developed in capability and performance in the last decade, facilitated by increasing computational resources in combination with the availability of language corpora, and driven by the demands of real-world applications in dictation, enquiry, indexing, and, increasingly, education. However, we are still in the early stages of applying speech technology within second language learning, and reactions from teachers and students are mixed [5]. Partly this is to do with pedagogical choices about how to use the technology to facilitate learning, but also there does seem to be real problems in how speech technology deals with accented speech. Speech recognition systems have problems in recognising the speech of second-language learners using acoustic models built from the speech of native speakers; evaluations of pronunciation similarity seem not to be well correlated with teacher judgements; and technological assessments do not always translate readily into advice that the learner can assimilate. In this paper, I would like to demonstrate some recent scientific advances in the way in which accented speech can be recognised, evaluated and manipulated which could improve the application of speech technology within language learning.

Our work at UCL on accent and speech technology has been to investigate fundamental issues about accent in general rather than second language accents in particular. So much of our experimental work has been based on studies of regional accents of English within the British Isles. However, I believe that the improvements in technology that are coming out of this work will also benefit applications in language learning: for example, through a richer approach to modelling the variability of phonological systems across speakers, or through a clearer separation in the acoustic signal of the influence of accent from the influence of speaker characteristics.

In section 2, I will describe some work in phonological adaptation in speech recognition that allows speech recognition systems to adapt to speakers not just in terms of phonetic quality but in terms of changes to phonological inventory and its use. In section 3, I will describe some work on accent recognition which explicitly differentiates between a speaker's accent and a speaker's voice. In section 4, I will describe some work that shows how accented speech can be manipulated to improve its intelligibility to native listeners. In each case I will give some suggestions for how these improvements in the underlying science could lead to improvements in the application of the technologies in language learning.

2. Recognition

The overall aim of our work in speech recognition is to improve the performance of automatic speech recognition systems on speakers of a known language but an unknown accent. Recognition results show that a mismatch between the accent of the test speaker and the accents of the training speakers can lead to significantly poorer recognition performance [3]. We believe that a large part of the problem is related to the overly simplistic assumptions about phonological and phonetic variety that are built in to recognisers.

In contemporary speech recognition, the dominant method for modelling the acoustic variability of speech within a language is to use a linear segmented phonological representation to structure the acoustic models of words. Typically a small set of phonological units ("phones") are chosen, often comprising just the phoneme set plus units representing silence and non-speech sounds. Word pronunciations are then commonly represented in the dictionary as just single phone sequences. Even when multiple pronunciations are used it is rare that these be assigned either prior probabilities (based on their frequency of occurrence) or conditional probabilities (based on the contexts in which they are found). Each phone unit is then associated with a number of statistical acoustic models, which capture the range of acoustic forms of those phones as realised by a large number of training speakers reading some known sentences. The acoustical models capture both variability in context and variability across speakers according to the structure imposed by a single phonological system.

There are two main ways in which such systems deal with speaker variety: (i) to sort speakers into one of a few groups, and to switch acoustic model sets according to the group, and (ii) to adapt the acoustic model sets towards the speaker's pronunciation using productions of a few known adaptation sentences. The first approach could be used to adapt to accent, but is most commonly only used to adapt to the speaker's sex, with different models for male and female speakers. The reason is that to use the first approach to adapt to accent would require enough labelled training material for each accent, a mechanism to assign speakers to a accent group, and an understanding of what accent groups are required. Not all of these are available for every accent of interest. However, some progress has been made in this direction for large accent groups [2].

Thus the dominant method for coping with accent is just the second technique which shifts the means of the statistical distributions of the acoustic models towards the measured means of an individual speaker. Significantly, such an approach assumes that the speaker's variation in pronunciation does not extend to the pronunciation dictionary or to the inventory of phones. In fact this makes adaptation an inadequate way of dealing with accent variation (in for example regional varieties of English within the UK) where changes in inventory (e.g. merging of vowel categories) or changes in phonological description (e.g. rhoticity) are commonplace. Neither is adaptation a good approach for dealing with foreign learners, since again their problems are not just of phonetic realisation, but also of contrast and pronunciation choice, with likely interference from the phonological and phonetic forms of their first language.

What is required are approaches to adaptation of the pronunciation dictionary itself. The naïve approach to include all possible pronunciations of every word in the dictionary can actually make matters worse, and give a lower level of recognition performance than a dictionary with just one entry per word. This is because multiple pronunciations per word reduces the average distance between words. When recognising an utterance there is no constraint that the set of pronunciations chosen for the words form a coherent and possible accent.

The obvious alternative, then, would be to build accent-specific dictionaries and combine these with a method for recognising which dictionary is most suitable for a particular speaker. However this approach has problems too, firstly because it assumes that phonetic knowledge about every accent is available, and secondly because it assumes that speakers can be indeed be put into one of a few categories.

An alternative has been proposed by my student Michael Tjalve [6], and he has shown that it gives superior performance to either approach. It is also intellectually more satisfying because it relates not to accent but to recurring pronunciation patterns that operate across groups of words in the lexicon. In the new approach, pronunciations of words in the lexicon are labelled as demonstrating the action of particular accent features. Thus the pronunciation of "mark" as [mɑːrk] would be labelled as obeying a rhotic rule, while the pronunciation of "butter" as [bʌɾə] would be labelled as obeying a flapping rule. During adaptation, the activity of each of a small list of possible rules are measured using a specially configured recogniser that performs a forced recognition of some adaptation sentences. From the set of active rules, a dictionary can be constructed containing only one pronunciation per word that best fits the single speaker, we call this an *idiodictionary*. The text box below gives some more detail of one experiment.

Experiment 1. Recognition using an Idiodictionary

Hypothesis: idiodictionaries built from accent features would be better adapted to a speaker than an accent dictionary chosen by accent recognition.

Data: Training set: 69,615 utterances from 247 speakers of British English. Adaptation set: 25 phonetically-rich sentences from 158 speakers of 14 different accents chosen from the Accents of British English corpus. Test set: 100 short sentences from the same 158 speakers.

Tools: Hidden Markov model recogniser using triphone contexts, Unisyn pronunciation dictionaries from 5 major British English accents [7].

Conditions: Baseline: sentence recognition accuracy using standard English pronunciation dictionary. Accent dictionary: accuracy using the best accent-specific dictionary. Idiodictionary: accuracy using individual idiodictionaries; these are made by choosing the most frequent of six accent features exhibited by each speaker within the adaptation sentences. and then constructing a specific pronunciation dictionary that implements those features.

Results:

Condition	Sentence Recognition Rate (%)
Baseline	71.8
Best Accent Dictionary	74.2
Idiodictionary	77.3

Conclusions: The use of an accent specific dictionary does indeed improve performance, with a reduction in sentence error rate by 8.5% over the baseline. However this assumes a perfect mechanism for assigning dictionaries to speakers, so even this small reduction may not be realisable in practice. However the use of idiodictionaries reduced the error rate by 19.5% over the baseline, and does not need a mechanism to allocate a speaker to an accent group.

What are called accent "features" here, and which are used to model phonological variation across accents, could also be called systematic pronunciation errors within a language learning system. For example, pronunciations of English that fail to differentiate "red" from "led" could be described by an accent feature that merges /l/ and /r/ in a group of words. When an idiodictionary is built by finding which accent features describe a learner's pronunciation best, what we are actually doing is making an analysis of the differences between the speaker and the standard phonological system of the target accent. The accent features could even be selected for specific L1-L2 pairs based on knowledge of common problems.

It is also worth pointing out that construction of an idiodictionary is complementary to normal adaptation of acoustic models, and preliminary work suggests that the improvements from dictionary adaptation and model adaptation are additive. This separation of phonological variety from phonetic variety could also be exploited in computer aided pronunciation teaching, where the learner can be told which phonological choices were incorrect and separately what phonetic realisations are in need of adjustment. However, it is still necessary to improve the way phonetic quality differences are judged by the technology, and this is the topic of the next section.

3. Measurement

Accurate analysis and recognition of accent, as well as judgement of pronunciation quality, demands a sensitivity to the phonetic patterns used by a speaker independently from the characteristics which relate to his or her individual vocal anatomy and physiology. Approaches to accent recognition and pronunciation measurement built on speech recognition technology fail to do this since they are based on a spectral analysis of the speech sounds which confound both kinds of information [2]. Indeed, studies have shown that the biggest single contributing factor to the acoustic distance between speakers is actually their sex, not their accent [3]. This mixing of speaker and accent information leads to an insensitivity to small differences in pronunciation, and in turn this leads to mistaken views about accent variation, and to poor quality evaluations in computer aided pronunciation teaching.

In contrast, experimental phonetic accounts of accent tend to use vowel formant frequency features which have the advantage that they can be normalised using the range of formant frequency values available to the speaker (e.g. conversion from hertz to z-scores [1]).

However formant frequencies are a relatively crude measure of vowel quality only, and may not be robustly estimated from the speech signal.

What is required is a means to use the robust spectral-envelope features for the analysis of a speaker's accent in a way that is insensitive to a speaker's own vocal characteristics. The ACCDIST metric [4], developed at UCL, shows one way in which this may be achieved. ACCDIST compares pronunciation *systems* across speakers rather than the acoustic quality of the speech itself. A model of the pronunciation system for a speaker is found by measuring the similarity between his or her different phone realisations, and a correlation between pronunciation systems across speakers then provides a measure of accent similarity.

A conventional pattern recognition approach to assigning an unknown speaker to an accent group would be to select a set of features from a number of training speakers and to calculate the mean values these features take for each accent. Linear Discriminant Analysis (LDA) then investigates how members of each accent group typically vary with respect to the mean. The accent means and the pooled variance can then be used to determine the most likely accent group of an unknown speaker. For example, the average spectral envelopes of a set of vowels are measured from training sentences from known speakers of a group of accents, then the accent of an unknown speaker is identified by comparing that speaker's vowels against the accent means.

A major problem with this approach is that average vowel spectra vary with the speaker's vocal tract size as well as with accent, thus speakers of the same accent may still have rather different spectra. The solution in the ACCDIST metric is to use the relative similarity of vowels within a speaker's pronunciation system as the features for recognition, rather than the absolute quality of the vowels themselves. Thus the table of distances between the vowels produced by a speaker is used to characterise the vowel "map" used by a speaker for a set of known words. Different accents will have different maps, so the maps themselves can be used to identify accents. A typical experiment is described below.

Experiment 2. Accent Recognition with ACCDIST

Hypothesis: Accent recognition using spectral features will be influenced by speaker type. Normalised features help reduce sensitivity to speaker type, but better accent recognition performance can be obtained by comparing pronunciation systems rather acoustic forms.

Data: 20 short sentences from each of 10 male and 10 female speakers from each of 14 regional accent areas of the British Isles. Automatic phonetic alignment allows the identification of the quality of about 100 vowels from each speaker. The vowels are either analysed in terms of spectral envelope features (MFCC) or in terms of formant frequencies. The formant frequencies can be normalised using the mean and variance of their values within each speaker. The ACCDIST metric calculates a pronunciation map for each speaker.

Tools: Linear Discriminant Analysis is used to compute the distance from each speaker to the means of the accent groups formed by all the other speakers. Pronunciation maps are compared by simple correlation.

Conditions: Spectral features: LDA based on spectral envelope features; Formant frequency: LDA based on raw formant frequencies; Normalised formant frequency: LDA based on z-scores of formant frequencies; ACCDIST: accent distances computed with the ACCDIST metric. Each metric is also evaluated using three gender conditions: Same sex: when speakers are only compared to other speakers of the same sex; Any sex: when speakers are compared to both sexes; and Other sex: when speakers are only compared to speakers of a different sex.

Results: Percentage correct accent group assignment for held-out speaker:

Condition	Same Sex	Any Sex	Other Sex
Spectral envelope	82.1	78.8	55.5
Formant frequencies	85.4	82.1	47.4
Normalised formant frequencies	84.3	86.9	74.5
ACCDIST	85.8	92.3	84.3

Conclusions: The results show that accent recognition based on the use of spectral envelope features or un-normalised formant frequencies is indeed sensitive to speaker type. We can see significant increases in performance when we limit recognition to the same sex, and significant drops in performance when we force recognition to the wrong sex. The normalisation of formant frequencies to the typical range used by the speaker helps a great deal, but there is still a significant fall in performance between the same-sex and the other-sex condition. This shows that speaker type is still an influencing factor even within one gender. In contrast the ACCDIST metric, which compares vowel maps not vowel quality across speakers, shows no significant drop in performance caused by the gender of the speakers, in addition it has the overall highest performance on the accent recognition task.

The ACCDIST metric seems a promising approach to accent recognition, but more than that, it seems to provide a means for comparing pronunciations of utterances across speakers. The results show not only good accent recognition performance, but also an independence to speaker type. ACCDIST could be extended to deal with consonantal and timing differences, and so form the basis for a pronunciation similarity score between native and learner utterances.

Other work on ACCDIST at UCL has been to cluster speakers into accent groups from the bottom up. This could lead to new data-driven approaches to the description of accent. We have also investigated how the correlations between the pronunciation systems could be studied with respect to the most significant differences. By finding which vowels contribute most to any fall in correlation between speakers, we can identify which vowels are most important in defining accent differences. We might then use this as the basis for feedback to a second language learner, or even demonstrate what the improved pronunciation would be like in their own voice, as the next section describes.

4. Manipulation

It is not only speech recognition technology that has developed in recent years. Technologies for manipulating and synthesizing speech have also improved considerably: from systems for voice conversion and prosody manipulation to unit selection synthesis and multi-lingual text-to-speech systems. It is now perhaps time to look at how these technologies for building and manipulating speech signals could be applied to accented speech. For example it is possible to envisage systems which could take a recording of a known phrase by a speaker and modify the speaker's accent using knowledge of the acoustic form and relationships between accents. So a recording of an actor could be modified to change their accent, or a recording of a second language learner could be modified to demonstrate a more native-like production.

Systems for modifying speech include: unit-selection synthesis, prosody manipulation and voice conversion. Unit selection synthesis rearranges the segmental content of recorded speech to make new utterances, prosody manipulation changes the pitch and timing of an utterance, while voice conversion changes the speaker identity of an utterance.

In unit-selection synthesis, a speaker records a large number of known sentences and these are analysed and labelled to identify the speaker's realisation of phonological units in context. These labelled signal components may then be combined to create new phrases by choosing units that fit together well. This has become the dominant method for signal generation in modern text-to-speech synthesis systems.

Prosody manipulation systems can change the pitch and timing of a recording by manipulation of the waveform itself. Techniques for manipulation are now of good quality, and providing the size of the changes are small, cause few processing artefacts.

Voice conversion systems map the spectral characteristics of one voice to another, such that a recording in one voice can be spoken out in another voice. Typically these are built using statistical signal processing techniques which are trained using parallel aligned corpora of the two speakers speaking the same sentences. Although such systems were originally designed to change speaker within an accent, some researchers have investigated using similar approaches to change the speaker's accent [8]. However the challenge here is to make pronunciation changes which preserve the speaker's identity. Before this can be addressed, we first need to assess which aspects of pronunciation need changing to convert an accent.

At UCL we are interested in the general question about the intelligibility of one accent by a listener of a different accent. One way to investigate this is to manipulate accented speech and discover the effect of the manipulations on listeners.

My student Kayoko Yanagisawa has been investigating which aspects of English-accented Japanese cause most problems for native Japanese listeners. She has been able to show that computer manipulation of prosody can indeed make English-accented Japanese significantly more intelligible. See the experiment described below for more details.

Experiment 3 – Requirements for Automated Accent Correction

Hypothesis: broadly we can divide the differences between English-accented Japanese and native Japanese in terms of segmental quality, pitch and timing. If we were to build a system to "correct" English-accented Japanese, would it be more important to change the phonetic quality, the pitch or the timing? We gauge importance in terms of how intelligible the manipulated speech would be to native listeners.

Data: intelligibility word lists in Japanese are read by a mono-lingual English speaker (working from a romanised respelling) and by a matched native Japanese speaker.

Tools: the recorded words are phonetically annotated and analysed for pitch and timing. This provides us with three data sets in each language representing the segmental quality component (Q), the pitch component (P), and the timing component (T) for each word. PSOLA prosody manipulation is used to change the pitch and timing of the Japanese recording to the English and vice versa.

Conditions: There are 8 conditions: $Q_EP_ET_E$, $Q_EP_ET_J$, $Q_EP_JT_E$, $Q_EP_JT_J$, $Q_JP_ET_E$, $Q_JP_ET_J$, $Q_JP_JT_E$, $Q_JP_JT_J$,. The words are played to 8 native Japanese listeners in a balanced factorial design. The recordings are mixed with pink noise at 3dB SNR to prevent ceiling effects.

Results: The table below shows mean word recognition rate pooled over the Quality, Pitch and Timing conditions:

Condition	English-accented (%)	Native-Japanese (%)
Quality	39.5	46.8
Pitch	33.3	53.1
Timing	41.8	44.5

Conclusions: As expected, correcting the English-accented recordings in terms of either quality, pitch or timing shows an increase in recognition rate by native listeners. However the increase in performance caused by changes in segmental quality or by changes in timing are small and not significant in statistical terms. The correction of pitch, did however make a significant improvement in recognition rate. This is undoubtedly due to the lexical role of pitch in Japanese that is not found in English.

Although this was just a pilot, this experiment showed that audio manipulation of accented speech can be used to increase its intelligibility to native listeners. The increase occurred even though the manipulation itself introduced small but inevitable processing artefacts into the signal. This results suggests that accent correction by computer is indeed possible: it really does address phonetic deficiencies in foreign-accented speech.

It is therefore worth investigating whether the accent manipulation of audio recordings would also have some value within second language learning. A particular role could be in a better means of providing feedback to learners about pronunciation errors. Improved pronunciations could be played back to the student in his or her own voice. It would be expected that these would be easier for the learner to assimilate than feedback in the voice of the teacher.

5. Conclusions

The application of speech technology to language learning is still at an early stage, and presents new challenges particularly with regard to accented speech. Research in the way in which the technology deals with accent in general will lead to a better understanding of accent variation, to improvements in the performance of the technology on accented speech, and to more successful applications within second language learning.

6. Acknowledgements

I would like to thank Michael Tjalve and Kayoko Yanagisawa for their contributions to this article. The work on ACCDIST was greatly influenced by related work by Nobuaki Minematsu.

7. References

[1] Adank, P., Smits, R., van Hout, R., "A comparison of vowel normalization procedures for language variation research", JASA 116 (5) 2004.

[2] Arslan, L., Hansen, J., "Language Accent Classification in American English", Speech Communication 18, 353-367, 1996.

[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] Huckvale, M., "ACCDIST: a metric for comparing speakers' accents", Proc. International Conference on Spoken Language Processing, Jeju, Korea, October 2004.

[5] Neri, A., Cucchiarini, C., Strik, W., "Automatic Speech Recognition for second language learning: how and why it actually works", 15th ICPhS Barcelona, 2003, p1157.

[6] Tjalve, M., Huckvale, M., "Pronunciation variation modelling using accent features", Proc. EuroSpeech 2005, Lisbon, Portugal.

[7] Unisyn lexicon: http://www.cstr.ed.ac.uk/projects/unisyn/

[8] Yan Q, Vaseghi S, “Analysis, Modelling and Synthesis of Formants of British, American and Australian Accents”, Proc ICASSP, 2003