ADAPTATION TO REGIONAL ACCENTS IN AUTOMATIC SPEECH RECOGNITION

WILLIAM BARRY

University of Cambridge
Linguistics Department
Cambridge CB3 9DA
United Kingdom

ABSTRACT

A method of speaker-adaptive speech recognition is presented in which systemic differences are exploited to identify the speaker's gross regional accent: A small number of "calibration" sentences are spoken by the prospective user. Intra-sentence comparisons are made of selected vowels differing between dialects in their systemic value, and the speaker is scored on strength of adherence to one of four gross regional accents. The regional accent decision and the numerical data derived from the analysis of the calibration sentences are used to modify values in the vowel reference tables.

REGIONAL ACCENT DIFFERENCES

Speaker-independent automatic speech recognition requires a solution to the problem of regional accent differences. The accent has first to be identified, and then the reference values used in the recognition process have to be adapted towards the particular accent.

Differences between accents exist at various levels of description. Firstly, there may be differences in the phoneme inventory. For example, many speakers of Northern British English do not distinguish the vowels in "look" and "luck", or "put" and "putt"; many Scottish speakers have the same quality vowel in "good" and "food".Secondly, even in those parts of the vowel system that have equivalent phonemic oppositions, the lexical distribution of phonemes may differ. This may be due to different historical development in a large number of words such as /æ/ in "path", "grass", etc. in American and Northern British English while Southern British English has /a:/. Alternatively, there may be isolated incidences, such as "tomato", which has /ei/ in American and /a:/ in British English. Thirdly, regional accents differ in the phonetic quality of functionally equivalent phonemes. For example, although Southern and Northern British can both be said to have a distinctive contrast between the vowels in "cat" and "cart", that distinction is not carried to the same extent by the same phonetic properties. The qualitative difference between /æ/ and /a:/ in some areas of Northern England is very small, the distinction relying almost totally on the length difference; in Southern British the qualitative difference is very noticeable.

SYNTAGMATIC COMPARISON

These differences can be exploited for recognition purposes by comparing the acoustic characteristics of selected vowels within a known text. Two known words may contain different quality vowels in one dialect and the same quality vowel in another. Whether the reason is a difference in inventory, lexical distribution, or just a difference in the phonetic relationship of functionally equivalent phonemes, analysis will provide evidence for or against a particular regional accent. This principle of text-internal or 'syntagmatic' comparison has an obvious advantage over comparison with any external template values. The relational values are obtained from the individual's own realisational framework, avoiding the problem of having to normalise for non-dialectal inter-speaker differences.

DELIMITATION OF REGIONAL ACCENT

Although regional accent variation is strictly speaking non-discrete, both in geographical terms moving from one area to another, and in sociological terms within a given area, some people <u>are</u> categorisable according to their geographical background. Four gross accent areas were selected for differentiation: Southern Standard British (SSB), Northern British (NB), Scottish (Scot), and General American (USA). The differences within these regions may well be regarded by some (particularly those who live in them) as being at least as great as the differences between them. They do, however, constitute regional accents which are readily recognised in everyday speech communication by linguistically naive persons, and must therefore be considered to have some identity. General American, in particular, is not a natural regional accent associated with one geographical area. It is a standardised accent, roughly equivalent in the United States to SSB in Britain.

PHONOLOGICAL DIFFERENCES

These gross accent areas offer phonological and phonetic differences in their vowel systems which can be exploited for identification purposes. On the basis of evidence collated in Wells [3], the phonemic differences, tabulated in Table 1, are theoretically sufficient to distinguish between them. Words (in capitals) are used to represent the phonemic oppositions because differences in the lexical distribution of phonemes and variation in phonetic quality make the choice of one symbol rather than another confusing. The word labels used are those employed by Wells [3, p.127ff.].

Table 1. Primary vowel comparisons for dialect separation.
+ : phonemic opposition
– : no opposition exists

	SSB	NB	Scot	USA
TRAP-BATH	+	–	–	–
FOOT-STRUT	+	–	+	+
FOOT-GOOSE	+	+	–	+

These three oppositions differentiate the British accents more clearly than the American accent. Three further vowel comparisons provide additional dialectal definition for American English. They also provide additional characterisation of Scottish. These may be considered "secondary" comparisons (Table 2).

Table 2. Secondary vowel comparisons for dialect separation

	SSB	NB	Scot	USA
LOT-CLOTH	–	–	–	(+)
LOT-THOUGHT	+	+	–	(+)
LOT-PALM	(+)	+	+	–

The bracketed indication of an opposition for USA in the LOT-CLOTH and LOT-THOUGHT oppositions are a necessary acknowledgement of differences within North America. Although a distinctive contrast is claimed for both of them in General American, there are many speakers who make no contrast. The bracketed opposition for SSB LOT-PALM is an indication that the vowel quality distinction is unreliable; the opposition relies more strongly on the length difference of the two vowels.

Another type of difference provides useful additional sub-grouping, namely the incidence of the long monophthongs /a:, ɔ:/. In so-called 'rhotic' dialects, that is in our Scottish and USA speakers, they do not occur in words spelled with an <r> following the vowel. In addition, of course, these dialects do not have /ɜ:/, which occurs in words such as "bird", "hurt" or "heard", nor the centering diphthongs and triphthongs (as in "here, hair, tour, hire, hour", etc.).

CALIBRATION SENTENCES

The accent classifier operates on sentences containing the word classes given in Tables 1 and 2. Practical usefulness to a speaker-independent recognition system requires that the sentences satisfy two conflicting criteria: they have to be as short as possible yet provide all the vowel comparisons, in stressed position, necessary for differentiating the target accents, if possible more than once for greater reliability. Ideally, to provide a representative picture of a speaker's vowel space, they should also contain at least one token of the vowels not required for the comparisons.

The following four sentences satisfy all these requirements:

1. After tea father fed the cat.
2. Father hid that awful cart at the top of the park.
3. Father cooked two of the puddings in butter.
4. Father bought a lot of cloth.

In sentence 1 we have a difference in distribution. Although both SSB and USA have an /æ/–/a:/ distinction, /æ/ occurs in many words in USA which have /a:/ in SSB. Thus, when comparing "after", "father", and "cat", an American speaker will have a the same vowel quality in "after" and "cat" and a different quality in "father"; the SSB speaker will have the same quality in "after" and "father" and a different quality in "cat".

In sentence 2 the difference between rhotic /ar/ in "cart" and "park" and the non-rhotic /a:/ in "father" will signal a Scottish and an American accent. SSB and NB have a non-rhotic /a:/ in all three words. In addition, less difference between "awful", and "top" than between "father" and "top" would be evidence for a Scottish or a Northern British speaker.

Sentence 3 provides an example of complete neutralisation. Northern British, in contrast to SSB has no distinction between the vowels in "cook" or "pudding" and "butter". A Scottish speaker, on the other hand, will have the same quality vowel in "cook", "pudding" and "two", a strongly fronted, close, rounded vowel.

In sentence 4, minimal differences between the vowels in "bought", "lot", and "cloth" would signal Scottish; similarity between "bought" and "cloth", with both words differing from "lot" would indicate USA; the same quality in "lot" and "cloth", and a large difference between each of these words and "bought" would be evidence for SSB.

In addition, the sentences also contain stressed words with the vowels /i:/, /ɪ/, and /e/, completing the inventory of stressed pure vowels, except for /ɜ:/. This is important if the accent identifier is to be used for anything more than pure diagnosis.

As comparison conditions are fulfilled, points are allocated for or against particular dialect categories. Positive and negative scoring aids differentiation. In some cases, fulfillment of a condition is evidence for one regional accent but strong evidence against another. For example, in sentence 1, a large difference in quality between "after" and "father" coupled with similarity between "after" and "cat" is strong evidence for an American accent and against Southern British. In other cases, an accent category is indifferent to nonfulfillment, and no negative points are allocated. For example, in sentence 3, Northern British will score positively, and Southern British negatively if "cook", "pudding", and "butter" have similar vowel qualities, but the USA score will be unaffected, due to a tendency for many American speakers to centralise both /ʊ/ and /ʌ/. Classification of a given speaker is based on the maximum accumulated score gained by any regional accent.

An obvious weakness in the practical application of the accent identifier is its present use of rigid and relatively gross criteria. Speakers with strongly modified accents can still be detected as non-standard by the human listener by means of other, perhaps finer regional features, which the accent identifier ignores. For example there are consonant and prosodic features which differ widely from one accent to another which have not yet been incorporated. At the moment, the only step towards differentiation of the degree of adherence to a regional accent is obtained from the continuous record of mark allocation, which can track the features which may deviate from the overall regional accent decision.

ANALYSIS PROCEDURE

Analysis is carried out in two steps. The first is a dynamic programming procedure to locate the vowels in the input sentences to be analysed. The second step is the comparison procedure itself.

The dynamic alignment uses a symmetric DP matching algorithm [2] operating after endpoint location on a combined measure of average amplitude per 20ms frame (normalised to compensate for differences in recording level) and zero-crossing count. After alignment, the analysis frames of the input sentence correspond to the frames in the reference sentence containing predefined comparison points; the comparison points are located manually with a speech-signal editor approximately one third through the selected vowels and the values stored.

The comparison procedure is an LPC-based, three-formant Euclidean distance calculated on an auditory (equivalent rectangular bandwidth) scale. The use of auditory scaling has the advantage of reducing the effect of F3 variation while giving very low F3 in rhotic vowels sufficient weight to influence the difference value. The formants are obtained by second derivative peak-picking, and cleaned by applying combinatorial constraints derived from phonetic theory. The constraints can be made extremely powerful by the fact that the vowels are known. It is also possible to inhibit individual vowel comparisons if a plausible formant structure is not found, thus avoiding totally spurious accent judgements.

In general, the formant analysis has proved very reliable, only falling down when the endpoint location of the input sentence, prior to the dynamic alignment procedure, fails due to extraneous noise. The use of a combined zero-crossing + amplitude measure for endpoint location provides considerable resistance to non-periodic disturbance.

ADAPTATION TO ACCENT

Accent identification itself is only the first step towards better recognition of non-standard speech. Adaptation is the necessary second stage. Part of this is possible on the basis of independent regional speech data, part depends on data on individual speakers gained from the calibration sentences during the identification process.

Independent vowel formant data for the regional accents have been collected from /hVd/ syllables. These provide regional group average values against which individual regional speakers' vowels can be matched. However, direct formant-to-formant matching assumes that, apart from vocal tract length differences, speakers differ only in regional accent. However, there can be other differences in long-term articulatory patterning [1] within accent groups resulting in differences in the exploitation of F1 and F2 space. In impressionistic terms this can be seen as the tendency of some speakers to speak without much jaw movement, or without much forward-and-back tongue movement. Adaptation to these differences is also possible on the basis of formant data gathered during the accent identification process.

Firstly, a group average F1/F2 "centroid" value is calculated from, the average vowel values, each vowel in the regional system being related to the centroid by an F1 and an F2 factor. In addition, the maximum and minimum F1 and F2 values give the group F1 and F2 "dispersion" values. Individual "centroid" and "dispersion" values are calculated from the calibration-sentence data. Adapted vowel target values are calculated by applying the regional group vowel factors to the individual centroid values, using the F1 and F2 dispersion factors (= individual dispersion / group dispersion) to stretch or squeeze the vowel space in the F1 or F2 dimension.

SUMMARY AND DISCUSSION

The accent classifier is conceived as a first stage of a complex front-end component in a speaker-independent speech recogniser. The correct classification of a speaker's accent is essential information which will be passed up the system, enabling, for example, the subsequent front-end sub-components to adapt to the speaker. It may also be needed to trigger a particular subsection of phonological rules, and to direct accent-dependent lexical access. However, more than just the accent decision can be exploited in the speaker adaptation process, which can be envisaged basically as a 'mapping' of the acoustic space in which the particular speaker produces his vowels. Analysis data from the calibration sentences provides an economical basis for this mapping procedure.

Problems not addressed by the approach described here are, male/female speaker normalisation, and modification for degree of regional adherence. Progress in the latter depends to a large extent on long-term data obtained from the accent classifier revealing which oppositions most frequently differentiate the speakers. As data accumulates, statistical evaluation will determine the relative frequency of occurrence of particular regional features. The hierarchy thus obtained can be used to specify degrees of regional accent and associate them with particular vowel features.

REFERENCES

[1] Nolan, F. J., 1983 The Phonetic Bases of Speaker Recognition. Cambridge: Cambridge University Press.

[2] Sakoe, H. and Chiba, S., 1973 Comparative study of DP pattern matching techniques. Speech Research Group, Acoust. Soc. Japan Report S73-22.

[3] Wells, J.C., 1982 Accents of English. Cambridge: Cambridge University Press.

ACKNOWLEDGEMENTS

This paper is based on work carried out as part of Alvey research project MMI/069 on Automatic Speech Recognition, funded via SERC grant GR/D/42405 under the collaboration of STL, Cambridge University, and the MRC Applied Psychology Unit. I am grateful to colleagues on the project and in the Linguistic Department for discussion and helpful criticism.