



## LINGUISTIC VERSUS PERSONAL VARIATION IN SPEECH RECOGNITION

Francis Nolan\*

### ABSTRACT

This paper concerns the adaptation of automatic speech recognisers to new speakers. Existing recognisers, in their training and adaptation, treat between-speaker variation essentially as acoustic 'noise' and ignore structuring which originates at higher levels, caused for instance by accent differences. If a large-vocabulary recogniser is to cope efficiently with a realistic range of speakers it will have to incorporate linguistic knowledge about accents. A solution to the problem of disentangling accentual and personal characteristics of new voices is outlined, and the subsequent adaptation of different components of a recogniser is discussed.

### INTRODUCTION

In most existing speech recognisers the process of adapting the device to a new speaker treats between-speaker variation as if it were an arbitrary, unstructured acoustic distortion. A simple template-matching recogniser for isolated words, which the new speaker trains by offering tokens of each word in its vocabulary, would be indifferent to whether the 'new speaker' is in fact a previous user speaking through a sock, a different individual speaking the equivalent of the words in a different language, or a group of people each responsible for speaking one particular word in the vocabulary.

A more sophisticated adaptation procedure involves extracting an acoustic characterisation from a limited sample of the new speaker's speech, and on the basis of that, performing a general transformation of the acoustic reference material in the recogniser.

This has the advantage that only a subset of a large vocabulary has to be actively trained; but if it is operating at the level of whole words it fails in two respects to acknowledge structure in between-speaker variation. On the one hand, assuming speakers with exactly the same dialect, the differences in their acoustic speech signals are structured by (at least partially) systematic differences in their vocal physiology. On the other hand, assuming speakers with different accents, their acoustic differences are additionally structured by their divergent phonological systems.

The cost of ignoring this latter, phonological, structuring can be seen with a simple example. Assume a recogniser whose reference material is based on American English (AmEng) pronunciations. If it is trained on a subset of its lexicon by a speaker of Southern British English (SBE) it will more or less successfully adapt to the global acoustic properties of

---

\*Department of Linguistics, Cambridge University, Sidgwick Avenue, Cambridge CB3 9DA, U.K.

the new speaker; but it will totally fail to adjust to the fact that a substantial minority of the set of words which in AmEng have the same vowel as cat have, in SBE, a different vowel - the same one as in calm. Thus in AmEng pass rhymes with gas; in SBE it doesn't. Conversely harm and calm rhyme in SBE, but not in most AmEng. Such cases are not isolated; and they apply between all major English accents. Scottish English, for instance, rhymes shoot and put, but not horse and hoarse.

Accents, then, don't just pronounce a particular sound differently (though this is important); they also restructure the lexicon in major ways. In a small-vocabulary recogniser, requiring each word to be trained, such restructuring will be problematic only if two required words turn out to be homophonous in the new accent. But as soon as a non-trivial adaptation process is attempted, involving generalisation from a small number of words to a large vocabulary, accent differences will cause problems.

These can be solved only when the recognition process explicitly reflects phonological structure - most probably by incorporating analysis into phoneme-length segments or features - and can draw on linguistic knowledge about the relationship of accents.

Two further advantages accrue, in principle, from basing recognition on phoneme-length segments. Firstly, and most importantly, it allows unrestricted addition of entries to the lexicon without adding acoustic reference data - since any new word added to the lexicon will be made up from the existing inventory of segments. Secondly, it opens the way to phonetically-informed adaptation to personal voice quality (PVQ). As is well known, the speech-acoustic values yielded by (for instance) a larger and a smaller vocal tract are not straightforwardly related. Because of the differential dependency of different resonances on areas of the vocal tract as it changes shape, their frequency relationship across two speakers cannot be modelled by a simple scaling factor. Transformation of data to a pseudo-auditory, or to a logarithmic, frequency scale may improve the effectiveness of a simple scaling factor; but it is likely that some sensitivity to the phonetic class of segments will be required of the scaling procedure. Such a phonetically-informed scaling, as part of speaker-adaptation, is clearly most feasible in a segment-based recogniser.

It is against this background that the Cambridge-STL Alvey project\* has been exploring the explicit inference of, and adaptation to, accent and PVQ characteristics in speech recognition.

#### INFERENCE OF SPEAKER CHARACTERISTICS

When a new speaker addresses it a speech recogniser is potentially faced with three unknowns - the linguistic content of the speech, the accent with which it is spoken, and the personal voice quality of the speaker. None of these three kinds of information presents itself discretely; all are mapped

---

\*Alvey MMI/069, SERC grant GR/D/42405

onto the same acoustic dimensions.

In the present project we assume, as a heuristic limitation, a co-operative speaker who will speak a number of pre-determined calibration phrases designed to contain phonetic material crucial to identifying the speaker's accent. Unfortunately acoustic differences between the input speech and any stored reference material are ambiguous - they could result from accent, or from PVQ. Particular acoustic values for a vowel segment, for instance, might reflect an accent feature; or the size of the speaker's vocal tract.

The solution we have adopted to this problem is to identify the accent by means of comparisons internal to the input speech, thus sidestepping the problem of PVQ differences between input and reference material. Barry (ref 1) describes the technique in more detail; briefly, the process is as follows. Accent-diagnostic sounds are located in the input calibration phrases. Location at present involves time aligning (using DTW) the input with a stored reference, but other strategies are possible. Next, spectral differences are computed - either as whole-spectrum distances or in terms of extracted formant frequencies - between relevant sounds. A scoring procedure incorporating phonetic knowledge is then applied to assess from the pattern of differences which of four major accent types best accounts for the new speaker. Thus, for instance, if the vowel of bath is spectrally further from that of calm than that of gas, this would weigh against Southern British English.

Once a decision has been made about accent, it becomes meaningful to compare the input calibration utterances with stored (accent-specific) phonetic material to assess personal voice quality characteristics. As noted above, it has long been established that the relationship across different speakers of spectral features expressed on a Hertz scale is rather complex. More recently it has been suggested that the relationship can be made simpler and more uniform by simulating the auditory system's transformation of acoustic data. Thus as a preliminary to detailed modelling of the new speaker's PVQ we have explored the usefulness in this respect of Bark and ERB transformations of spectral data (ref 2).

#### ADAPTATION TO SPEAKER CHARACTERISTICS

In a large-vocabulary recogniser based on phonetic segments adaptation to accent will have to be provided for at several levels. Firstly, the lexicon will have to allow for the differences in structure imposed on it by the phonologies of different accents. This could be achieved by giving each lexical entry multiple phonological forms - one for each accent. Such a solution would entail an unnecessary degree of redundancy. Instead, we have chosen, where there is a substantially regular phonological alternation between accents, to use 'cover symbols' in lexical entries. In this way pass might be represented as /pAs/, reflecting the fact that the vowel phoneme of this word switches in different accents in a way which that of gas /gAs/ does not.

Then, secondly, there is accent-specific activation of rules. The form /pAs/ will be interpreted by rule to have the /æ/ of gas if the accent is AmEng, but the /a:/ of calm if it is Southern British.\* The rules in question are parsing rules which take from the recogniser 'front end' a (fully- or partially-specified) phonetic string and access a set of possible word strings from the lexicon. Hoequist (ref 3) gives an account of the implementation of such rules. In addition to their accent-specific expansion of lexical entries they also have to accommodate accent-specific behaviour in phonological processes such as segment deletion and reduction.

Even at the level of phonetic feature extraction, processing will have to be informed by knowledge of the speaker's accent. It may be crucial in one accent to distinguish phonetic types which are of no consequence in another; and the distribution of equivalent types in phonetic space may differ considerably. At this level, too, adaptation to PVQ needs to be applied if the phonetic types are to be recognised reliably for different speakers.

Speech recognisers for English and most other languages can be expected to encounter speakers with significantly different accents, as well as personal voice qualities. To lump these two kinds of variation together in a purely acoustic adaptation procedure is unlikely to succeed, because it ignores regularities in the two kinds of variation which can only be predicted from phonological, and acoustic-phonetic, knowledge, respectively. This paper has discussed the framework of a current attempt to incorporate such knowledge.

#### REFERENCES

1. W.J. Barry, Adaptation to regional accents in automatic speaker recognition. Proc. 11th Int. Cong. Phonetic Sci. (Tallinn), paper 24.3 (1987).
2. D.H. Deterding, Use of the ERB scale in peripheral auditory processing for vowel identification. Proc. 11th Int. Cong. Phonetic Sci. (Tallinn), paper 24.4 (1987).
3. C. Hoequist, Phonological rules and speech recognition. (This volume) (1987).

---

\*In cases where an accent-specific alternation is reliably conditioned by phonetic environment it may be unnecessary to introduce cover symbols in the lexicon; but frequently exceptions to general rules (such as gas, maths) mean that alternations have to be coded in the lexicon.