AUTOMATIC PHONETIC FEATURE LABELLING OF CONTINUOUS SPEECH

M.A. HUCKVALE, I.S. HOWARD, W.J. BARRY

Department of Phonetics and Linguistics, University College London

ABSTRACT

This paper updates our previous work on the automatic phonetic feature analysis of speech. Previously we have described how a bank of feature-detectors can be used as a front-end to traditional speech recognition pattern matching algorithms, with increased performance in speaker-independent isolated word recognition over purely acoustic front-ends. In this paper we extend the feature analysis to continuous speech: describing the labelling methodology and the additional classification performance of neural network classifiers over the Bayes normal classifier.

1. INTRODUCTION

In many speaker-independent continuous-speech recognition systems (e.g. 2,3), the speech signal is first analysed into a small number of spectral coefficients which are used to construct statistical models of the realisations of sub-word units of speech. These systems explicitly mix together speech from a number of speakers in constructing these models, since it appears that recognition performance is more affected by undertraining (under-estimating the observation variances) than by the added confusion of mixed speakers (4).

However it is well known that different phonological entities have overlapping acoustic realisations when multiple speakers are analysed (e.g. 5); that is to say knowledge of speaker characteristics affects the phonological interpretation of acoustic evidence.

Future recognition systems must acknowledge speaker differences at all levels in the speech decoding hierarchy; it will become necessary for the recognition system to have knowledge of speaker types and be able to adapt its recognition processes accordingly. A great weakness of existing systems is their reliance on statistical models of acoustic parameters which have a high degree of variability from speaker to speaker and from environment to environment. These statistical models need to be conditioned by speaker and environmental parameters, but there has been no mechanism by which this might be done.

At University College London, we have been pursuing the possibility of alternative front-ends for speech recognition systems that are sensitive to these issues: they normalise some speaker characteristics and could be made responsive to environmental conditions.

In this paper, we describe our hypothesis that a phonetic feature-based front-end based on a non-linear transformation of spectral parameters will offer increased speech recognition performance (section 2). We describe a corpus of data and the acoustic-phonetic annotations used for training and testing a feature analysis (section 3). We describe the construction and performance of a number of different feature detectors (section 4).

2. FEATURE ANALYSIS

To establish a representation of a signal that has been normalised for speaker characteristics requires a representation that make explicit the phonetic dimensions of the signal - that is the dimensions which are exploited to convey phonological contrasts in speech. The complete range of these contrasts is not yet established, and there are many subtle aspects of a signal that convey phonological information. However we have so far only been concerned with gross aspects of the acoustic-phonetic mapping, and to exploit simple dimensions/features such as voicing, nasality, frication, vowel quality, stop bursts and glides (note: with these terms we refer to acoustic properties, which unfortunately have to be named using phonological terminology). These are robust properties of the signal which are known to have phonological interpretation.

Our approach should be contrasted with phonetically-motivated Vector Quantisation, such as that used in the IBM systems (6). Whereas in these systems, labels from a small phone inventory are attached
to short sequences of the signal, we aim to give a multi-dimensional, continuous analysis of the signal without explicit segmentation.

Our rationale for feature analysis is similar to (7), where a feature matrix derived from gaussian classifiers is used directly in lexical access. We prefer to use a further level of pattern recognition on top of the feature analysis to recognise words (instead of segmentation and rules) - indeed we see our work as evolving from existing sub-word HMM systems.

Our central hypothesis is that a parameter set which makes explicit the phonetic dimensions of the signal is more appropriate than the direct use of spectra even for existing recognition algorithms. Furthermore it opens the possibility for recognition based on prior knowledge of the phonology of the language.

With this and previous papers, we have set out to test our hypothesis. Our basic approach has been to compare the performance of recognition algorithms working directly on a set of acoustic parameters and also via a non-linear transform of the identical parameters to a set of phonetically-motivated acoustic-phonetic features, See Figure 1.

In (1) we were able to show that in an isolated word speaker-independent recognition experiment, the use of the phonetic transform gave significantly better recognition results. However, this experiment had two weaknesses: (i) It used a simple dynamic programming matching algorithm with a Euclidean distance metric for the features. (However, we have recently repeated this experiment using Hidden Markov Models for the words and obtained equivalent results, these will be reported in (8)).

(ii) The features were designed for and trained on the vocabulary words.

In this paper we set out to derive a set of broad features trained on continuous speech, that would form the basis for a more complete feature-set appropriate as a general front-end for speech recognition systems.

3. SPEECH MATERIAL

The speech material used to train the feature detectors was the four English continuous speech passages on the ESPRIT ‘SAM’ project CD-ROM: EUROMO. These were analysed with a 19-channel filterbank to the specification of (9), with energies low-pass filtered at 50Hz and sampled at 100Hz.

The passages were then manually annotated at an acoustic-phonetic level using an inventory of 186 symbols. The aim of the labelling was to delimit distinct acoustic segments within sections of the speech signal identifiable as phonemic units. For example, a common variety of the phoneme /k/ consists of (1) a closure, (2) a distinct burst, (3) a section of prevocalic aspiration. Acoustic segment labelling locates and specifies each of these constituents. However, even greater differentiation is often possible: the closure, for example, might easily be partially voiced following a vowel, and that periodicity in the signal would be labelled as a separate acoustic segment.

The labelling convention uses a basic structure of a digraph-symbol; the first part specifies the phonemic unit, and the second part gives the subcategorising property. The above example of /k/ would therefore be labelled as:

kv + kc + kb + kf
(voiced voiceless burst aspiration)
(closure closure noise frication)

The properties differentiated within the phonemic units are:

Voiced - v (pv, gv, lv etc.)
Voiceless - c (closure)  (tc, kc, bc, etc.)
Noise + voice - a (ia, ma, la, etc.)
Nasality - n (an, En, etc.)
Diphthongal - d (ad, Ud, etc.)

The only exception to the digraph convention is a trigraph in the case of the second element in a diphthong, which is marked by a ‘2’. Thus the diphthong /AI/ would appear as:

ad + Id2

The annotations were then mapped into a feature matrix using a look-up table. The features were:

FASP aspirated and voiced section
FWEAK weak and strong frication
FSTRONG strong frication
WWEAK weak and strong voicing
VSTRONG strong voicing
FRONTV front and central vowel quality
BACKV back and central vowel quality
CLOSEV close and central vowel quality
OPENV open and central vowel quality
GLIDE diphthongal glide, plus /j,r,w/
NAS nasal
BURST stop burst
Entries from the look-up table were of the form <annotation><list of marked features>, e.g.:

ia  FRONTV CLOSEV FASP VWEAK
if  FWEAK
in  FRONTV CLOSEV VWEAK VSTRONG NAS
iv  FRONTV CLOSEV VWEAK VSTRONG

The four passages together comprise 500 seconds of speech, labelled with 7500 annotations. Pattern vectors were constructed from three adjacent filterbank frames and the binary feature value appropriate for the centre frame. This provided 40,000 pattern vectors for training each feature and 7,500 pattern vectors for testing.

4. FEATURE DETECTORS

For each of the 12 features listed above, for each of the speakers independently, two types of detector were trained and tested: a Bayes classifier for normal patterns ('gaussian' classifier), and a neural network classifier using the multi-layer perceptron algorithm with back error-propagation/steepest-descent training (10). Each perceptron had 57 input units (3 frames of 19 channels) 10 hidden units and a single output unit.

The gaussian classifier was trained by determining the mean and diagonal covariance matrix for the on and off states of each feature, and tested using the Bayes decision rule and equal penalties for misses and false-alarms. Each perceptron was trained by presenting it with 30 passes over the training set, updating the weights every 4000 pattern vectors. The perceptron output was simply thresholded to determine output class.

Classification performance on the test data is shown in Table 1. Column 1 displays the name of the feature, while column 2 gives the percentage of test vectors that were marked with that feature. Columns 3 and 4 give the percentage hit rate and percentage false-alarm rate for the gaussian classifier. Column 5 shows the MLP classification performance with a threshold chosen to give equivalent false-alarm rate as the gaussian classifier. In every case, the MLP gives better performance. Column 6 shows the MLP classification performance with a threshold chosen such that the miss-rate is equal to the false-alarm rate (i.e. the equal-error point).

Figure 2 shows sample outputs of the MLP classifier on the fragment 'Germany's decision' extracted from the test data.

5. CONCLUSIONS

For nine out of the twelve features, frame labelling rates were 80% or better (at the equal-error point), this is quite good performance given that it includes classification errors at boundaries between annotations. The features FASP, GLIDE and BURST, however had poor performance, and this must be put down to the inhomogenous nature of these classes. Examination of the causes of error will aid us in designing replacement features.

The generation of a robust set of features is only the first part of the story. We now move on to exploiting these and other features for word recognition. The feature analysis allows us to generate predictive models for the range of realisations of an arbitrary word given only its phonetic transcription. Exploiting transcription-based models in our HMM work we are already finding that start-up word models with a feature-based front-end have better recognition performance that equivalently constructed start-up models using the filterbank directly.

6. REFERENCES

(1) I S Howard, M A Ruckvale, 'Training feature detectors for use in automatic speech recognition', 7th FASE symposium, SPEECH-88, Edinburgh, September, 1988
(2) Kai-Pu Lee, Hsiao-Ruen Hon, 'Large-vocabulary speaker-independent continuous speech recognition using HMM', ICASSP-88, p123.
(5) G E Peterson, H L Barney, 'Control methods used in a study of the vowels', JASA 24 (1952) p175.
Table 1 - Performance of Feature Detectors

<table>
<thead>
<tr>
<th>Feature</th>
<th>%Frames</th>
<th>Gaussian %Hit</th>
<th>Gaussian %False</th>
<th>Network %MLP1</th>
<th>Network %MLP2</th>
</tr>
</thead>
<tbody>
<tr>
<td>FASP</td>
<td>5</td>
<td>76.7</td>
<td>64.6</td>
<td>90.6</td>
<td>63.4</td>
</tr>
<tr>
<td>FWEAK</td>
<td>18</td>
<td>94.2</td>
<td>36.4</td>
<td>96.8</td>
<td>84.7</td>
</tr>
<tr>
<td>FSTRONG</td>
<td>9</td>
<td>91.7</td>
<td>39.8</td>
<td>93.1</td>
<td>89.0</td>
</tr>
<tr>
<td>VSTRONG</td>
<td>50</td>
<td>87.4</td>
<td>5.3</td>
<td>95.4</td>
<td>95.3</td>
</tr>
<tr>
<td>VWEAK</td>
<td>52</td>
<td>78.0</td>
<td>7.8</td>
<td>95.0</td>
<td>93.5</td>
</tr>
<tr>
<td>VSTRONG</td>
<td>36</td>
<td>71.7</td>
<td>9.9</td>
<td>81.5</td>
<td>87.2</td>
</tr>
<tr>
<td>BACKV</td>
<td>17</td>
<td>78.5</td>
<td>17.1</td>
<td>84.4</td>
<td>83.8</td>
</tr>
<tr>
<td>OPENV</td>
<td>21</td>
<td>82.2</td>
<td>12.4</td>
<td>91.5</td>
<td>89.9</td>
</tr>
<tr>
<td>CLOSEV</td>
<td>31</td>
<td>69.9</td>
<td>17.5</td>
<td>89.1</td>
<td>86.2</td>
</tr>
<tr>
<td>GLIDE</td>
<td>5</td>
<td>60.1</td>
<td>24.1</td>
<td>66.8</td>
<td>68.9</td>
</tr>
<tr>
<td>NAS</td>
<td>15</td>
<td>85.7</td>
<td>23.9</td>
<td>94.4</td>
<td>88.9</td>
</tr>
<tr>
<td>BURST</td>
<td>1</td>
<td>85.7</td>
<td>43.8</td>
<td>89.2</td>
<td>72.3</td>
</tr>
</tbody>
</table>

Figure 2 - Example Feature Analysis

Figure 1 - Recognition Paradigm

speech signal \rightarrow acoustic parameters \rightarrow phonetic transform \rightarrow recognition

file=/mod/eurom/ea.part speaker=ea token="Germany's decision ...

Figure 2 - Example Feature Analysis

Time (s) 6.0 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8 6.9

Amplitude

Annotated features

Feature analysis

BURST
NAS
GLIDE
CLOSEV
OPENV
BACKV
FRONTV
VSTRONG
VWEAK
FWEAK
FSTRONG
FASP

Time (s) 6.0 6.1 6.2 6.3 6.4 6.5 6.6 6.7 6.8 6.9