Proceedings of the Institute of Acoustics

PHONETIC FEATURE ANALYSIS FOR A MONOSYLLABIC WORD RECOGNITION TASK

M Huckvale & I Howard

Department of Phonetics and Linguistics, University College London, Gower Street London WC1E 6BT

1. INTRODUCTION

> "the interconversion of phone and sound is an integral part of language and its underlying physiology." [1].

In our research work we have been attempting to construct a continuous phonetic feature description of speech signals in which segmentation is performed at equal time intervals (10 or 20ms) and where the value of each feature in an interval represents the probability of the signal exhibiting a particular phonetic property (for a justification, see [2]). We have constructed this type of representation using perceptron networks that are trained to perform non-linear transformations of the signal [3]. We have previously demonstrated the utility of the approach in a simple speaker-independent digit recognition task [4].

Our hope is for an automated procedure for performing phonetic analysis on all speech signals regardless of source. This could become an essential component of the computer-speech systems of the future, in the same way as the equivalent human procedure is part of human language 'physiology' as Mattingly and Liberman suggest in the quote above.

Although such a universal procedure is long distant, we suggest that a start may be made on simple sub-sets of speech signals:- single speakerrs, single environments, restricted linguistic form. If we can establish a scientific procedure for determining the phonetic transforms for these sub-sets independently, then in future we might consider meta-level procedures for matching transforms to speech signals in general.

In this paper, we make a first attempt at the design and implementation of an experimental paradigm for the determination of a set of phonetic transforms from a database of speech material. Section 2. below describes the experimental paradigm in general terms, while section 3. describes our specific implementation for a monosyllabic word recognition task. Sections 4. and 5. give results for a simple vowel and consonant recognition subset.

2. PHONETIC TRANSFORM DERIVATION PARADIGM

The essential characteristics of the experimental paradigm are (i) the phonetic transformation is embedded in a phonological recognition task, and (ii) there is feedback from phonological confusions to the design of the phonetic transformation.

PHONETIC FEATURE ANALYSIS

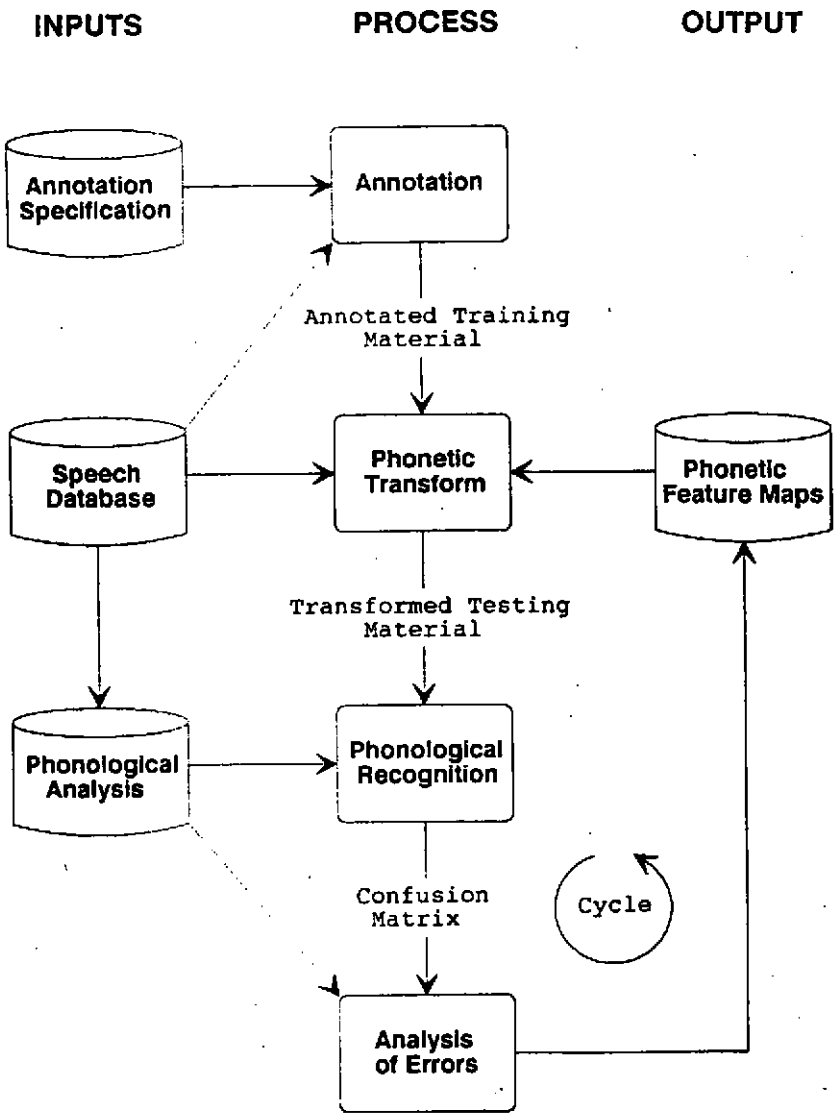

Fig 1. Schematic of an experimental paradigm for the determination of a phonetic feature transform to represent some speech material. The output is the feature transform specification rather than the transform itself.

PHONETIC FEATURE ANALYSIS

The outputs of the paradigm are not the transforms themselves, since these will be speaker and environment dependent, but rather the specifications for the transforms: the 'feature maps'. These assume that we can annotate signals reliably and build transforms to specification using some adaptive procedure, and so can define a transform according to the required relationship between annotated region and feature output. The feature maps say such things as: regions annotated with '-m-', '-n-', '-l-' should have the VOICE feature high; or regions annotated with 'p-burst' should have the ONSET feature high for 10ms. Fig 1. gives a schematic view of the paradigm, with the feature maps seen on the right as output. There are three 'inputs' to the paradigm:

1) Annotation Specification: a formal statement of the procedures to follow to associate regions of the speech signal with labels. These labels need not be tied to phonological units, and would normally be chosen to simplify the annotation process (to ensure reliability of annotation). Since the annotations must later be used to identify regions which have different phonetic properties, the annotations must at least be specific enough to identify different phonetic regions. The annotation specification for an experiment would normally consist of a set of labels and a set of criteria for associating those labels with regions of the signal.
2) Speech Database: a collection of material chosen to represent some well-defined subset of speech signals. The material must be reasonably homogeneous and self-consistent: small vocabularies or single speakers or single environments. The paradigm aims to produce a phonetic feature transformation appropriate for this material.
3) Phonological Analysis: the phonological task used to assess the effectiveness of the phonetic transformation requires some given phonological labels for the speech database material. These labels are chosen to represent at minimum the phonological diversity of the database material; they need not be a complete or parsimonious set for the language.

The paradigm has 4 procedures with which it functions:

1) Annotation to Specification: given the speech material and the annotation specification it is also necessary to have a formal procedure by which one can be assured that the material is annotated to specification. It is important that the annotations are used reliably; if consistency is difficult, then the specification should be reconsidered.
2) Programmable Phonetic Transformation: a procedure for the construction of a transform from acoustic parameters to phonetic feature probabilities built from a feature map specification. A number of pattern recognition tools could be considered here, and feedback will be required that modifies the feature map to fit their capabilities.
3) Phonological Recognition: a recognition algorithm operating on the output of the phonetic transforms and attempting to select phonological labels for the speech material. This is used to establish whether the phonetic transforms maintain sufficient information for phonological choices to be made.
4) Error Analysis: tools for taking the results of the recognition to aid the making of improvements to the set of feature maps.

3. EXPERIMENTAL METHOD

Our current implementation of the experimental paradigm described above has the following components as inputs:

1) Annotation Specification: 125 labels have been selected to cover the acoustic-phonetic events in the speech material (below); these are simply related to a traditional articulatory phonetic transcription, but pragmatically extended to ease annotation of complex segments or smooth transitions (more information in [3]). An example section of speech signal has also been identified for each label.
2) Speech Material: a 1000 monosyllabic-word vocabulary has been selected to cover a large subset of syllable structure in English. 334 words have been arbitrarily chosen for training, 333 words for evaluation and 333 words for final testing. Attempt has been made to get maximum coverage out of dictionary words. The words were recorded by one speaker (MH) in an office environment with a close-talking microphone and automatic endpointing. Further details and recordings are available from the authors.
3) Phonological Analysis: each monosyllabic word was analysed as three segments: onset, nucleus, coda; with the consonant 'clusters' treated as single phonological entities. Thus the recognition task was to separately identify the initial consonant cluster, the vowel and the final consonant cluster.

The current implementation of the procedures was as follows:

1) Annotation to Specification: consistency was assured (at the expense of accuracy of time placement) by using a dynamic programming alignment procedure to align a specified annotation label sequence each to the utterances. The example annotated signal segments were used as a source dictionary from which an artificially-created utterance for the word could be aligned with the original.
2) Programmable Phonetic Transformation: we have continued to use the multi-layer perceptron algorithm in a supervised training procedure [3]. Input to each network was a 30 ms window of a 19-channel filterbank analysis of the speech, output was the required feature value. For training, the feature map identified whether the network should be high, low or indifferent to each of the possible annotated regions. There was one feature map and one network per feature. The networks were trained to try to achieve high fidelity between map and actual performance of the network. This occasionally required changing the map to better fit the performance of the network.
3) Phonological Recognition: the output of the feature transform networks was fed to a set of Hidden Markov Models (HMMs), one per phonological unit. These were simple chains of 5 or 7 states with no skips. Each observation vector was modelled with a set of Gaussian distributions with diagonal covariance. Initial segmentation and distributions were set up using the procedure of Bridle & Sedgwick [5]. HMMs were re-estimated until there was a less than 1% change in model likelihoods.
4) Error Analysis: at each cycle in the experiment, a hypothesised set of feature maps were specified and the transformation trained accordingly on the training data. Outputs of the transformation again on the training data were then used to train a set of phonological

models (for the initial consonants or the vowels or the final consonants). These were then tested on the evaluation data to give confusion matrix analyses. Two methods have been used to investigate the confusion matrices: Information Transfer Analysis, where phonological subsets are chosen to explore which features are currently being exploited; and Multi-Dimensional Scaling, where phonological subsets are determined a posteriori. Both techniques can lead to suggestions for modification of the feature set for a new cycle in the experiment.

4. VOWEL EXPERIMENT

The vowel experiment to be described below gives a simple demonstration of the current implementation of the paradigm. The vowel experiment uses the monophthong subset of the database for the phonological recognition procedure. Thus the task is to define a feature set which adequately discriminates the phonological labels: /i, I, e, &, V, A, O, Q, u and 3/.

To obtain a reference level of performance, a set of HMMs were trained directly on the 19-channel filterbank energies of the whole words. The recognition rate on the evaluation database was 53%; the confusion matrix is shown in Fig 3a.

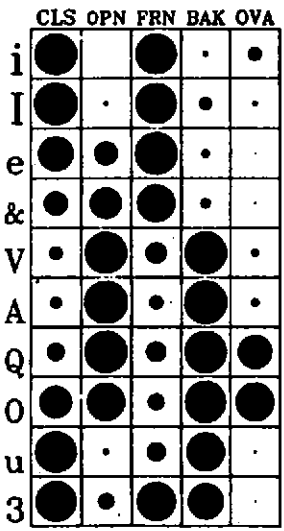

Fig 2. Vowel feature performance. The diameters of the circles represent the percentage of annotated region marked above threshold.

The following were chosen as an initial a priori set of features:

a) NUC: high when the speech signal is part of the syllable nucleus, low otherwise.
b) CLS: high when the vowel quality is for a 'close' vowel, low when an 'open' vowel, don't care otherwise.
c) OPN: high when the vowel quality is for an 'open' vowel, low when a 'close' vowel, don't care otherwise.
d) FRN: high when the vowel quality is for a 'front' vowel, low when a 'back' vowel, don't care otherwise.
e) BAK: high when the vowel quality is for a 'back' vowel, low when a 'front' vowel, don't care otherwise.

The selection of which parts of the signal represent front/back or high/low vowel quality was made in terms of the annotations. The networks were left to decide how to label the half-open and half-front vowels. The performance of the vowel quality features as a function of annotated region is shown diagrammatically in the first 4 columns of Fig 2.

The phonological unit recognition performance with these 5 features was 44%. The confusion matrix for this configuration is shown in Fig 3b. The most obvious conclusions to be drawn from an analysis of the confusions is the need for further separation between the open back vowels. Thus a 6th vowel feature was trained:

f) OVA: high when the vowel quality was /O/ or /Q/, low when /A/ or /V/, don't care otherwise.

The performance of this feature is shown diagrammatically in the 5th column of Fig 2. The phonological performance with these 6 features rose to 60%, slightly higher than the reference. The confusion matrix is shown in Fig 3c.

5. CONSONANT EXPERIMENT

This experiment looked at the monosyllabic words with single initial consonants, i.e. from the set /0, b, d, g, p, t, k, m, n, l, r, w, j, dZ, tS, f, s, S, T, v, z, D and h/. The pattern vectors for testing were generated from the beginning of the recording for each word to half way through the vowel (as determined by the annotation alignment procedure).

Reference performance was again obtained by training a set of HMMs directly on the vocoder energies. The result was 44% correct from the 23 phonological categories. To simplify the analysis, the tokens and models were pooled into the broader manner categories: 0, VSTOP, UVSTOP, VFRIC, UVFRIC, NASAL and LIQUID. The broad category recognition rate was 67%; confusion matrix in Fig 4a.

The initial set of features for the consonant recognition task was:

a) ENV: Amplitude envelope feature. The MLP configuration to implement this feature was constructed by hand.
b) NUC: high when the speech signal is part of the syllable nucleus, low otherwise.
c) VOI: high when the speech signal has periodic excitation, low otherwise.
d) FRC: high when the signal has aperiodic excitation, low otherwise.
e) NAS: high when the signal is nasalised, low for other voiced consonants, don't care otherwise.

Recognition rate on the 23 phonological categories was very poor: 17%, however these features provide no place information, so a fairer comparison would be with the 7 broad categories, with a recognition rate of 57%; the confusion matrix is shown in Fig 4b.

An analysis of the broad category confusions suggested that the primary sources of error were (i) liquids being mis-recognised as voiced fricatives and nasals, and (ii) voiced stops being mis-recognised as unvoiced fricatives. Two additional features were thus added:

f) LIQ: high when signal is syllable initial l-, r-, w-, j-; low for initial m- and n-, and voiced fricatives; don't care otherwise.

g) BUR: high for 10ms after a stop burst, low for voiceless fricatives, don't care otherwise.

With these 7 features, recognition rate went up to 20% for the 23 phonological categories and up to 65% for the broad categories, still slightly worse than the reference; confusion matrix in Fig 4c.

6. SUMMARY

In this paper we have outlined an experimental paradigm by which a phonetic feature specification may be derived from a speech database which may be used to develop a phonetic transform of speech signals to accomplish some phonological recognition task. We have also shown, in two simple experiments, that embedding the design of the transform in a recognition task allows us to hypothesize and test feature specifications, leading to an increase in performance. Whilst a system for the recognition of the vowels or initial consonants of isolated words may be of limited use, we hope the feature specifications will ultimately have more general importance.

7. ACKNOWLEDGEMENTS

M Huckvale was supported in this research by a fellowship from the Science and Engineering Research Council, I Howard was supported under a project grant from the Royal Signals and Radar Establishment.

8. REFERENCES

[1] I G MATTINGLY & A M LIBERMAN, 'The Speech Code and the Physiology of Language', in Information Processing in the Nervous System, ed K Leibovic, Springer Verlag 1969, p111.

[2] M A HUCKVALE, 'Exploiting Speech Knowledge in Neural Nets for Recognition', Speech Communication 9 (1990) pp1-13.

[3] M A HUCKVALE, I S HOWARD, W J BARRY, 'Automatic Phonetic Feature Labelling of Continuous Speech', European Conference on Speech Technology, Paris, September 1989.

[4] I S HOWARD & M A HUCKVALE, 'Two-level recognition of isolated digits using neural nets', IEE Conference on Artificial Neural Nets, London, October 1989.

[5] J S BRIDLE & N C SEDGWICK, 'A Method for Segmenting Acoustic Patterns, with Applications to Automatic Speech Recognition', ICASSP-77, pp656-659.

PHONETIC FEATURE ANALYSIS

Confusion Matrix

	i	I	e	&	V	A	O	Q	u	3
i	18	5	0	1	0	0	0	0	1	0
I	2	12	3	2	1	1	1	1	3	1
e	0	3	8	4	5	0	1	0	1	2
&	0	0	3	19	2	1	0	0	0	1
V	0	0	1	2	5	7	3	5	0	0
A	0	0	0	1	0	16	0	4	0	0
O	0	0	0	0	0	0	11	7	0	0
Q	0	0	2	0	3	4	8	9	0	0
u	0	1	0	2	1	0	2	0	16	0
3	0	0	0	3	5	0	0	0	2	6

Number of matches = 228
Recognition rate = 52.6%

Fig 3a. Vowel Reference

Confusion Matrix

	i	I	e	&	V	A	O	Q	u	3
i	6	16	3	0	0	0	0	0	0	0
I	1	10	13	1	0	0	1	1	0	0
e	0	0	14	8	0	0	2	0	0	0
&	0	1	2	23	0	0	0	0	0	0
V	0	0	0	0	0	0	0	23	0	0
A	0	0	0	0	0	0	1	20	0	0
O	0	0	0	0	0	0	6	12	0	0
Q	0	0	0	0	0	2	1	23	0	0
u	0	0	0	0	0	1	5	10	6	0
3	0	0	0	0	0	0	0	3	0	13

Number of matches = 228
Recognition rate = 44.3%

Fig 3b. Vowel 5 Features

Confusion Matrix

	i	I	e	&	V	A	O	Q	u	3
i	10	12	2	1	0	0	0	0	0	0
I	2	10	11	1	0	0	2	1	0	0
e	0	0	16	6	0	0	0	2	0	0
&	0	0	2	24	0	0	0	0	0	0
V	0	0	0	0	3	4	0	16	0	0
A	0	0	0	0	0	12	0	9	0	0
O	0	0	0	0	0	0	7	11	0	0
Q	0	0	0	0	2	0	0	24	0	0
u	0	0	0	0	1	0	2	0	19	0
3	0	0	1	0	2	0	0	2	0	11

Number of matches = 228
Recognition rate = 59.6%

Fig 3c. Vowel 6 Features

Confusion Matrix

	0	VSTOP	UVSTOP	VFRIC	UVFRIC	NASAL	LIQUID
0	0	0	4	1	1	2	3
VSTOP	0	16	9	2	5	1	5
UVSTOP	0	2	24	1	8	0	1
VFRIC	0	0	0	15	0	0	1
UVFRIC	0	1	2	1	33	1	0
NASAL	0	0	0	4	0	11	0
LIQUID	0	1	0	2	0	3	26

Number of matches = 186
Recognition rate = 67.2%

Fig 4a Consonant Reference

Confusion Matrix

	0	VSTOP	UVSTOP	VFRIC	UVFRIC	NASAL	LIQUID
0	0	6	3	1	0	0	1
VSTOP	1	13	9	0	11	0	4
UVSTOP	0	1	29	1	5	0	0
VFRIC	0	0	0	14	0	0	2
UVFRIC	1	1	3	1	32	0	0
NASAL	0	1	0	4	0	7	3
LIQUID	0	0	0	13	0	4	15

Number of matches = 186
Recognition rate = 59.1%

Fig 4b. Consonant 5 Features

Confusion Matrix

	0	VSTOP	UVSTOP	VFRIC	UVFRIC	NASAL	LIQUID
0	3	1	3	0	2	0	2
VSTOP	8	11	14	0	4	0	1
UVSTOP	0	1	34	0	1	0	0
VFRIC	0	0	0	12	0	0	4
UVFRIC	0	0	7	0	30	0	1
NASAL	0	1	0	3	0	8	3
LIQUID	0	0	0	7	0	3	22

Number of matches = 186
Recognition rate = 64.5%

Fig 4c. Consonant 7 Features