Proc. 15th Int. Congress of Phonetic Sciences, Stockholm, August 1995.

PHONETIC CHARACTERISATION AND LEXICAL ACCESS IN NON-SEGMENTAL SPEECH RECOGNITION

Mark Huckvale
University College London, London, U.K.

ABSTRACT

An isolated-word speech recognition system, built without the use of linear segments for acoustic modelling or lexical access, is justified, described and demonstrated. The system comprises phonetic feature analysis operating on four independent tiers, parallel phonotactic parsing, and lexical access based on a neural-network inspired lexicon structure. Performance is however still inferior to a baseline segmental system.

INTRODUCTION

This paper describes an attempt to bring together into a single operational system a selection of alternatives to the linear segmental approach to phonetic modelling and lexical access found in contemporary automatic speech recognition systems.

The most important departure from current architectures is the explicit separation of phonetics and phonology in the system. In the new system the role of the first is to characterise speech-specific elements of the sound signal, while the role of the second is to establish the functions of these elements in linguistic encoding. In contrast, current systems based on phones-in-context use linear phonological units to organise their acoustic models as well as for lexical access. Such systems have particular weaknesses, including (i) poor modelling of variation of acoustic realisation of phonological units in context, (ii) failure to model post-lexical phonetic variety because of the need for complex and arbitrary context-sensitive realisation rules, (iii) failure to exploit contextual variation as discriminative information, (iv) failure to use temporally extended information relevant to phonological identity, (v) failure to exploit prosodic structure in the signal. These weaknesses lead to systems which lack discriminative power, are unable to exploit known pronunciation variety in context or in accent, fail to extract the most from impoverished signals, and ignore the information and constraints available in the rhythm, stress and intonation of the speech.

On the other hand, linear phonological-unit based acoustic models provide a simple and computationally effective basis for recognition. There is a synergy between a linear phonological account and syntactic pattern recognition algorithms such as Hidden Markov Modelling (particularly the Viterbi decoding scheme). It has been said that in speech recognition good knowledge is of no use without good algorithms for applying it. Hidden Markov Modelling has been successful because it forms a *coherent* view of the acoustic to phonological mapping, rather than an accurate one.

Thus the challenge is to find effective procedures for the exploitation of more sophisticated models of speech.

DESIGN

In this section we justify the non-segmental recognition system described in the following section. More details may be found in [1].

Phonetic component

The role of the phonetic component in a non-segmental system is to model the range of variety of acoustic realisation of elemental phonetic characteristics. For each given characteristic at each time frame, the phonetic component supplies the probability that the element has been realised (by a given speaker in a given acoustic environment). By relaxing the

requirement that these characteristics need to be themselves phonological we can make this component more sensitive to sub-phonemic changes, to syllabic and prosodic structure. Although we can no longer exploit phonological sequence constraints we can still exploit phonetic constructional constraints that arise due to the fact that the signal was spoken. In the simplest model, the phonetic component operates on a number of *tiers* where the phonetic properties inside a tier are mutually exclusive, while properties across tiers are mutually independent. As we shall see this allows the use of a syntactic pattern recognition scheme to operate within a tier.

Lexical Access

From the phonetic characterisation of the signal it is necessary to explain the phonetic evidence as realisations of a sequence of words, subject to a number of constraints: (i) words occur strictly sequentially, (i.e. only one word is active at any one time), (ii) citation form phonetic structures of words are subject to a limited range of contextual modifications, (iii) word selection is guided by the task (vocabulary, syntax, etc.).

Since at this stage we do not have a phonological representation, all we can do is activate word hypotheses on the basis of the likelihood that they might have given rise to the phonetic evidence. Following the TRACE model of lexical access [2] we can see that each phonetic characteristic can feed 'activation' into the lexicon, (but in this case without an interposing phonemic layer). Given a tiered phonetic analysis, any single tier activates a number of possible word hypotheses. The initial activations of words need not be zero, since there may be prior evidence (from the task) for the likelihood of words.

Phonological categorisation

From the word activations (over time), it is necessary to determine the most likely word sequence. Unfortunately, what we have at the moment is essentially a whole-word template recognition system, and it is easy to show that such systems cannot be extended to large vocabularies without the exploitation of *phonological* knowledge. Each word has been activated on the basis of phonetic similarity with the input, but it is likely that some components of the word match better than other components. Thus the vowel of [pi] may match the input quite well, but the consonant may match badly. If each word has independent pronunciation models of phonetic realisation, it is possible that the vowel of [ti] might not match as well as the vowel of [pi]. Thus an input "'T" may be recognised as "P" because the vowel matches overcome the consonantal matches. The solution to this is to indicate that the vowel in [pi], i.e. /i/, is the same as the vowel /i/ in [ti]. With this constraint, the difference in the vowel scores is irrelevant and the consonantal match controls the outcome. This is *phonological* knowledge that must be specified in addition to the phonetic realisation of words.

One way of imposing these phonological constraints is to establish a set of phonological units above the words, which share activations between words which have similar phonological prescriptions. Thus an /i/ unit short-circuits activation between [pi] and [ti] to counteract exactly any difference due to independent models of the vowel.

ARCHITECTURE

The specific implementation of the non-segmental recognition architecture for an isolated word recognition task may be separated into: (i) multiple Phonetic feature components that deliver phonetic feature analyses of 30ms of speech signal, (ii) Phonotactic decoding components that deliver element sequence likelihoods for each tier, (iii) a Lexical access component that takes the

element sequence scores and delivers a word hypothesis using lexical and phonological information. More details may be found in [3].

Phonetic feature component

The phonetic feature component operates on four independent tiers, corresponding to multiple broad-class analyses of the signal.

In the *Excitation* tier, phonetic elements represent Silence (SIL), Voicing (VOI), Frication (FRC) and Mixed excitation (MIX). In the *Degree* tier, elements represent Oral closure (STP), Nasal (NAS), Fricative (FRC), Approximant (APP), Close vowel (CLS), Mid Vowel (MID) and Open vowel (OPN). In the *Position* tier, elements represent Labial (LAB), Dental (DEN), Alveolar excluding /s/ (ALV), /s/ frication (FRS), Front/Palatal (FRN), Central (CEN), Back (BAK), Velar (VEL) and Silence (SIL). In the *Strength* tier, the elements represent Burst (BUR), Aspiration (ASP), Other frication (FRC), Vocalic (VOW), Voiced plosive (VGP) and Silence (SIL).

These tiers together are sufficient to differentiate English words apart from short and long vowels at a single place (e.g. bit vs. beat) and dental and labiodental fricatives (e.g. thin vs. fin). Performance on elements for these contrasts is currently unsatisfactory.

For each tier, a Multi-Layer Perceptron (MLP) classifier was trained between a spectral representation of the signal and the target element classes. Each tier had its own MLP with 3x10ms frames representing 19 filterbank energies + overall energy (i.e. 60 parameters) as input and 1 output per element class. Each MLP had a single hidden layer of a size equal to three times the output layer size. The training data was 666 different monosyllabic words spoken by one speaker. There were approximately 83,000 training vectors.

Each training word was annotated and the element labels generated by rule using a mapping that took into account boundaries and the nature of adjoining segments. Training was performed using an adaptive back-propagation method firstly on the automatically generated element labels, and then, after realignment with the partially trained network, against realigned element labels.

Phonotactic decoding component

To generate an element sequence, a Viterbi decoding was performed on the MLP outputs for a tier over the whole duration of a word. See Figure 1. This process delivered a score for each phonotactically possible sequence in the test vocabulary for each tier. Over the 4 tiers there were 450 possible element sequences, but only the best scoring 50% in each tier were used for lexical access

Lexical access component

To identify the lexical item a network lexicon was used based on [1]. Here the phonetic input was provided by the element sequence scores; these then fed activations to the word units according to 'dictionary' pronunciations of the words. Thus words were only connected to element sequences expected in the citation pronunciation. To smooth activations across words, a level of phonological units were constructed above the word units, which channelled activation between words sharing similar phonological descriptions - in this experiment shared syllabic components. Thus word activations arose primarily from the phonetic input, but subsequently there was interaction and competition between words mediated by a set of phonological units. The most strongly activated word unit was chosen to be the recognised word.

RESULTS

For testing the architecture, 359 monosyllabic words, different to the training words, but spoken by the same

speaker were used. The raw recognition performance of the Phonetic feature analysis component was:

Tier	Frames correct
Excitation	91.6 %
Degree	82.8 %
Position	74.7 %
Strength	87.9 %

The raw recognition scores for the element sequences was:

Tier	Top 1	Top 5
Excitation	76.6 %	98.3 %
Degree	46.0 %	80.2 %
Position	23.4 %	52.9 %
Strength	44.0 %	88.6 %

For the feature-to-word activations alone, without the use of the phonological units for smoothing, the word recognition performance was 51%. Small amounts of phonological unit activation fed back to the word units improved recognition performance only slightly, to 53%. Performance is so weak primarily due to the poor performance of the Position tier.

Baseline recognition performance using a monophone HMM trained on the same material (and having approximately the same number of free parameters as the set of MLPs) was over 90%.

FURTHER INFORMATION

The author welcomes comments on M.Huckvale@ucl.ac.uk

REFERENCES

[1] Huckvale, M.A. (1990), "Exploiting Speech Knowledge in Neural Nets for Recognition", *Speech Communication*, p1.

[2] McClelland, J.E. & Elman, J.L. (1986), "Interactive Processes in Speech Perception: The TRACE Model", in *Parallel Distributed Processing*, ed Rumelhart & McClelland, MIT Press.

[3] Huckvale, M.A. (1994), "Word Recognition from Tiered Phonological Models", *Proc. IOA Conf. Speech and Hearing*, Windermere.

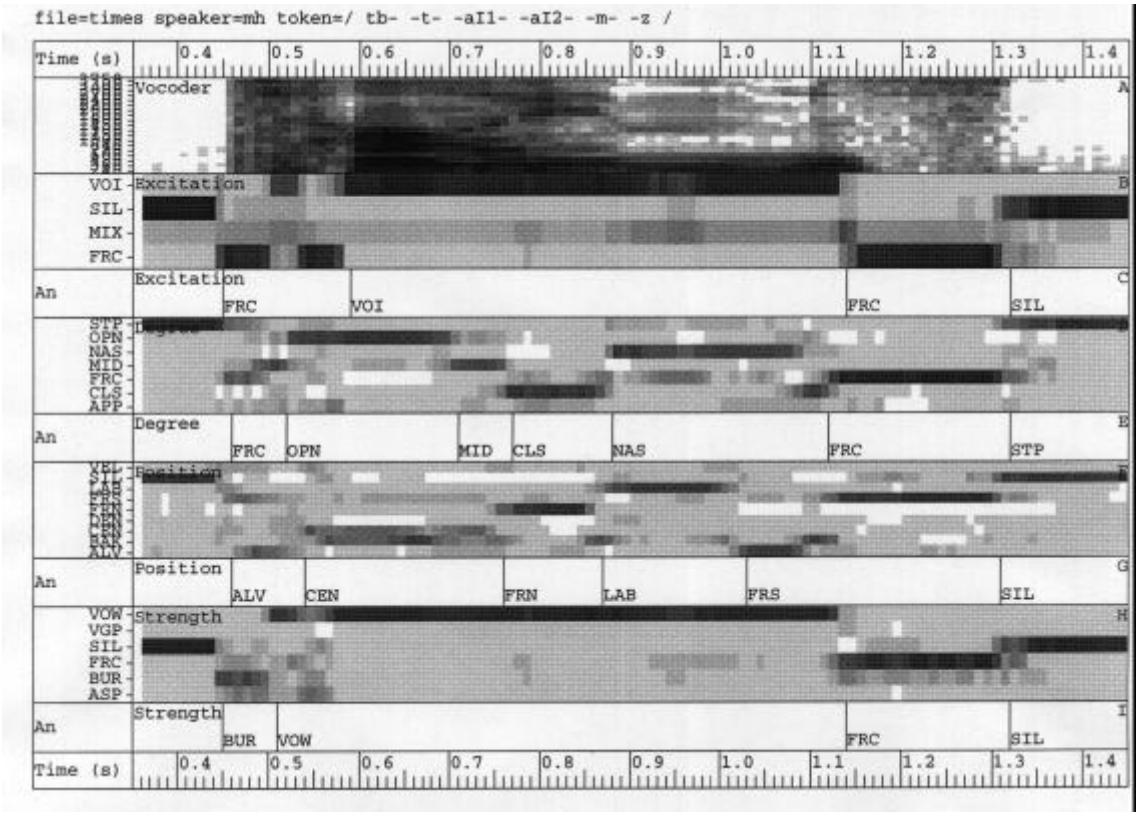

Figure 1. Tiered analysis of the test word 'times'.