From: Proc. Institute of Acoustics Conf. Speech and Hearing, 1994

WORD RECOGNITION FROM TIERED PHONOLOGICAL MODELS

M.A.Huckvale

Department of Phonetics and Linguistics, University College London

1. INTRODUCTION

This paper gives a snapshot of ongoing research into alternative phonological models of the acoustic structure of speech for exploitation in speech recognition systems. The basic approach is to recognise a number of fairly independent layers or 'tiers' of information about the phonetic content of the signal and to use these for lexical access. This contrasts with conventional approaches which use single layer segmental accounts of signal structure and a phonemic dictionary.

Over previous accounts of this research [1,2], this paper gives a different phonological argument for the use of multiple tiers, describes four new tiers, and reports word recognition performance using a new lexical access scheme.

There remain many difficulties with the tiered recognition approach, both conceptual and practical - these are outlined in a closing discussion.

2. PHONOLOGICAL BACKGROUND

On theoretical grounds, models of phonology which are closer to the phonetic reality of speech while providing sufficient lexical discrimination should provide a better basis for speech recognition systems. That is phonological units that have features more directly related to the acoustic-phonetic content of the speech signal should be more readily trained; and as long as the overall model provides enough power to discriminate vocabulary words, should provide increased recognition performance.

Contrast two linear phonological models in which one uses /H/ to stand for both the glottal fricative /h/ and the velar nasal /N/ (This is OK in phonological terms, because /h/ and /N/ are in complementary distribution - there are no minimal pairs), while the other uses two different units. We would expect the latter to provide a better basis for a speech recognition system, since we would hope that a pair of acoustic phone models [h] and [N] to be a better match to the signal than one joint model [H] covering both. (The 'within-class' variance would be smaller in the separate models, hence there should be better discrimination with other models).

However the phonetic fidelity of a phonological model cannot increase without limit: firstly because the amount of training material is always limited (and frequently

insufficient), so more units means poorer estimates; and secondly because more units mean more competition for explanation of a stretch of signal, so an increase in the fidelity of the units must be matched by an increase in the power of the sequential constraints (to give an example, there should be no points in the recognition grammar where different paths are solely reliant on differentiating /h/ and /N/). Thus the choice of the best phonological model cannot be made in isolation from the recognition task constraints (vocabulary, grammar and quantity of training material).

Let us re-consider why it is necessary to worry about the phonology at all for a given recognition task. For a small vocabulary task, a common approach is to treat each word as an independent phonological entity; then the number of units equals the number of vocabulary words, the lexicon is just the set of allowable phrases, and phonotactics is the same as syntax. However for a large vocabulary task, this 1:1 mapping is inadequate, not because it gives rise to too many units, but because it fails to model the structural variability of the signal. Independent models of the words 'pat' and 'bat' will contain different models of the variety of [t] which contribute differently to the distance between these models and an incoming "pat" or "bat"; whereas the differences in the acoustic form of [t] are *irrelevant* to the distinction. The power of the system to discriminate between 'pat' and 'bat' is weakened by not recognising the structural similarities between them: whereas one shared model of [t] would concentrate the metric on the syllable onset. Thus we need to consider how sounds function in words, not just which sounds make up a word.

Historically in ASR, the way to model the function of sounds in words is to make a bridge to a taxonomic phonological description - almost always a phonemic account. Thus 'pat' is presumed to have 3 units /p/, /a/ & /t/ because of lexical contrasts with the words 'bat', 'pit' and 'pack'. Notice how the '3-ness' comes from the words available in the lexicon, rather than from some acoustic or phonetic account. There is a large amount of human perceptual prejudice in this process, and quite a strong alphabetic influence.

However the introduction of a linear segmental description fits rather uncomfortably with the original intention, which was to find better models of the acoustic variability of words. Once linear units become the basis for acoustic models, we are essentially saying that the difference between 'pat' and 'bat' can be found in the difference between [p] and [b] independently of the vowel; whereas of course, the distinction is made in the context of the vowel - in the way the vowel (aka vocalic region) starts. The difference between 'pit' and 'bit' is still [p] vs [b] but with different acoustic events. Similarly, the difference between 'pat' and 'pit' is phonemically equivalent but acoustically different to the difference between 'wag' and 'wig'.

There is an interesting choice to be made at this point in the argument: should we (i) patch-up the phonemic account to allow for a range in acoustic realisations of phonemic units, or (ii) choose a different phonological account which is closer to the

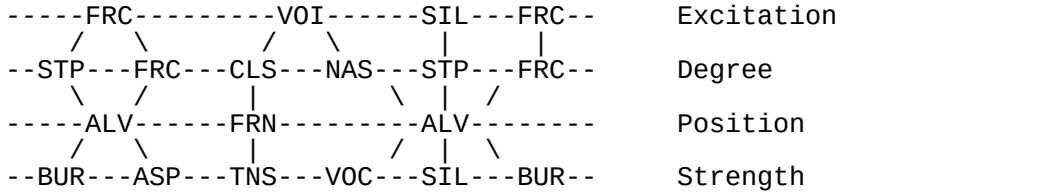
phonetic form. Since 99% of all ASR systems choose (i); and introduce 'phone-in-context' modelling; let us look at (ii).

What could be meant by an alternative phonological account? Such an account would still differentiate between words in the lexicon, but its units would be more directly related to the processes of production (and possibly perception). We would be looking for characteristics of words which would help to identify lexical items, but which had a smaller range of acoustic manifestations. The approach presented in this paper is to divide the phonological account into a number of layers, or 'tiers', whereby independent sources of variability are accounted for on different layers.

The selection of tiers in current accounts of non-linear phonology, e.g. [3], are rather too abstract for the purposes we seek here. A 'phonetic transparency' is required to have hope of robustly determining the presence of phonologically relevant characteristics in the signal. It seems appropriate to turn instead to studies of speech perception, for example studies which have analysed the perceptual confusions listeners make when higher level linguistic information is absent. A particular example are MDS (multi-dimensional scaling) studies, e.g. [4,5], in which perceptual confusions are transformed into a perceptual map with well-defined dimensions. These dimensions are strikingly similar to the old-fashioned Voice-Place-Manner labels of articulatory phonetics. Vowels have a clear two-dimensional pattern following the vowel quadrilateral, with the possibility of a third 'tenseness' dimension.

Let us take the phonetic dimensions voice, manner, place and tenseness; and relabel them acoustically as excitation, degree, position and strength to avoid confusing the phonological with the acoustic. These dimensions are interesting, not only because they have arisen experimentally, but because they relate fairly clearly to the articulation and acoustic properties of speech. Thus a phonological account that described how the speech changed along these dimensions within a word would be a phonology tied to an acoustic-phonetic description of the production of that word.

On such an account a prototypical phonological description of the word 'tinned' might look like:



i.e. the Excitation tier describes Frication-Voicing-Silence-Frication, the Degree tier describes Stop-Fricative-Close(Vowel)-Nasal-Stop-Fricative, the Position tier describes Alveolar-Front(Vowel)-Alveolar, and the Strength tier describes Burst-Aspiration-Tense(vowel)-Vocalic-Silence-Burst.

This phonological account differs in a number of ways from a linear phonemic account. Firstly it treats the dimensions as relatively independent sources of variation: a whispered word, a vowel that is less fronted, an aspirated final /d/, a missing burst - these are variations which are each isolated to a single tier. To an extent this independence comes from the link with the production system: tongue height, tongue frontness and phonation are independent. Secondly it puts weaker constraints on the time synchronisation of events in the different tiers; I have shown some vertical links which might describe some 'supporting' evidence which flows between tiers. It is not necessarily the case that tongue fronting follows the same time course as jaw opening or as changes in excitation. Thirdly, characteristics which are shared by adjoining segments are shown as single components (such as voicing in the vowel and nasal, or 'alveolar' placing of the nasal and final plosive). Fourthly an acoustic event may provide evidence for more than one phonetic unit (on more than one tier).

The tiered account is still capable of differentiating the words in the lexicon - it has the same power as a phonemic account - although it is more redundant than the phonemic account (it is not as parsimonious in features). To an extent this redundancy can be reduced through sequential constraints derived from an analysis of possible lexical configurations and applied on top of the phonological description. (This is no different to the phonemic account, where phonotactic restrictions are not implicit in the phoneme inventory). The redundancy also turns out to be necessary and useful for a practical implementation, which we now describe.

3. PATTERN RECOGNITION FRAMEWORK

3.1 Recognition Strategy

In previously reported work [1,2], both HMM (hidden Markov model) and MLP (multi-layer perceptron) syntactic pattern recognition schemes were used to perform tiered segmentation and recognition. The experiment below uses a single MLP per tier for feature extraction, followed by a viterbi decoder to provide a parsing for the tier. More details may be found in [1].

The use of an MLP with one output per element class within a tier places constraints on the best units to use within a tier. Every frame must be labelled with one (and only one) unit. This means that it should be impossible for two units to be active simultaneously within one tier. This means that, for example, that frication detection and burst detection must go in different tiers. This restriction actually fits in well with the original aim of the tiers, that they be independent dimensions of the speech signal. However the need to label every frame introduces some redundancy into the phonological account.

To aid in training the MLP, it is preferable to have fairly equal numbers of training

vectors for each of the element classes within a tier - or at least weight the outputs to their recognition importance. The tiers described below have been adjusted to even up the likelihood of the different classes, but more work is required.

3.2 Design of tiers

In the **Excitation** tier the Units are:

SIL	Silence
VOI	Voicing
FRC	Frication
MIX	Mixed Voicing and Frication

In the **Degree** tier the Units are:

STP	Oral closure
NAS	Oral closure + nasality
FRC	Fricative
APP	Approximant
CLS	Close vowel
MID	Mid vowel
OPN	Open vowel

In the **Position** tier the Units are:

LAB	Labial
DEN	Dental
ALV	Alveolar (excluding /s/)
FRS	/s/ frication
FRN	Front/Palatal
CEN	Central
BAK	Back vowel
VEL	Velar
SIL	Silence

In the **Strength** tier the Units are:

BUR	Burst
ASP	Aspiration
FRC	Other frication
VOW	Vocalic region
VGP	Voiced plosive
SIL	Silence

The strength tier should also have units for differentiating short and long vowels at a single place, and units for differentiating dental and labio-dental fricatives. At present performance on such units is unsatisfactory.

3.3 Process of training

Each tier has its own MLP with 3 or 5 frames of input (30ms or 50ms window) and one

output per class. Each has one hidden layer of a size equal to 3 times the output layer size. The training data is 1 repetition of 666 monosyllabic words spoken by one speaker, and analysed with a 19-channel filterbank + 1 channel energy. There are approximately 83,000 training vectors.

Each word has been phonetically annotated and the tier unit labels are generated by a mapping which takes into account the boundaries and the nature of adjoining segments. Training takes place using an adaptive error back-propagation algorithm for 20 passes over the training data. After a first MLP is trained, it is then used to realign the initial annotations, using a constrained viterbi decoding of the MLP outputs.

Recognition performance on the training data for each tier, with and without re-alignment for the 3x30 and the 5x20 input MLPs was:

% Frames Correct	3x20 Original	3x20 Realigned	5x20 Original	5x20 Realigned
Excitation	91.8	97.5	92.5	98.2
Degree	83.7	91.1	84.0	90.1
Position	75.3	83.5	74.6	82.4
Strength	81.2	90.2	87.8	93.6

On this basis, the 3x20 realigned MLPs were chosen for the recognition experiment.

4. RECOGNITION RESULTS

4.1 Test data

Testing was performed using one repetition of 359 monosyllabic words different from the training set. These were recorded at a different session by the same speaker, and analysed in the same way. See Figure 1.

Test performance was as follows:

% Frames Correct	3x20 Realigned
Excitation	91.6
Degree	82.8
Position	74.7

Strength	87.9

Frame correctness was of course measured against non-realigned annotations - so some performance drop was expected.

4.2 Tier-word recognition

To gauge word recognition adequacy of the raw MLP outputs, the 359 test words were collapsed into equivalence classes for each tier: so called 'tier-words'. Thus the words 'arms', 'is', 'wash' all have the Excitation tier word SIL-VOI-FRC-SIL. Using a viterbi decoding on the MLP outputs to select one of the allowable tier words (used in the 359 test words) gives the following tier-word recognition performance for various rank positions in the scores:

%Tier words contained	Top 1	Top 2	Top 5	Top 10	Top 25%
Excitation (12 possible)	76.6	91.9	98.3	98.6	96.1
Degree (150 possible)	46.0	60.4	80.2	87.2	93.9
Position (251 possible)	23.4	33.1	52.9	64.1	88.6
Strength (37 possible)	44.0	66.9	88.6	95.5	95.3

4.3 Word Recognition

Given the difference in performance of the tiers, it makes sense for word recognition to place more weight on the most robust tier first, then introduce the poorer tiers as required until a single lexical entry is found. This ensures that the weaker tiers are only used to resolve ambiguity when required. Since it is unlikely that the top scoring tier-words all belong to the same lexical candidate, some combination of scores across tiers is required to find the best scoring word.

In practice this was performed by firstly allocating a score to each lexical entry with a tier word appearing in the top 25% candidates in the Excitation tier. If this did not locate a single best-scoring candidate, then the Degree, Strength and Position tiers were brought in one by one. This appeared to make the best use of the recognition power of the independent tiers and the restrictions of a limited vocabulary.

For the same 359 test words, this process succeeds in identifying 51% correctly.

5. DISCUSSION

There remain many challenges for the tiered recognition scheme, both practical and conceptual.

Conceptually: (i) the tiers are too different in their complexity, performance and importance in lexical access, they need to be more homogeneous (articulatory place in particular is too complex a characteristic for a single tier); (ii) the lexicon needs to consist of more than one pronunciation per word and needs to model the pronunciation variability; (iii) the redundancy across tiers should either be reduced or be used more effectively in stabilising performance.

Practically: (i) it would be enormously preferable to have a 'parallel' viterbi decoding which provided the single best legal combination of tier-words rather than find the N best tier-words and combine the results; (ii) the recognition scheme should cope better with varying number of training vectors per class; (iii) the chosen acoustic representation should be different for the different tiers.

For further details of the data and procedures, contact the author on M.Huckvale@ucl.ac.uk.

6. REFERENCES

[1] M HUCKVALE, 'A comparison of neural-network and hidden-markov-model approaches to the tiered segmentation of speech', Proc. IOA Conf. Speech and Hearing, Windermere, (1992).
[2] M HUCKVALE, 'The benefits of tiered segmentation for the recognition of phonetic properties', Proc. EuroSpeech-93, Berlin, (1993).
[3] J DURAND, Generative and Non-Linear Phonology, esp. Chapter 6, Longman (1990).
[4] B RAKERD & R VERBRUGGE 'Linguistic and acoustic correlates of the perceptual structure found in an individual differences scaling study of vowels', JASA 77 p296, (1985).
[5] S SINGH, D WOODS, G BECKER, 'Perceptual structure of 22 prevocalic English consonants', JASA 52 p1698 (1972)

WORD RECOGNITION FROM TIERED PHONOLOGICAL MODELS

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