TIERED SEGMENTATION OF SPEECH: OPPORTUNITIES, METHODS, PROBLEMS AND CHALLENGES

Mark HUCKVALE

Abstract

This paper discusses the opportunities for automatic speech recognition systems provided by a multi-dimensional phonetic representation of speech signals. Through discussion of methods and an example implementation of tiered recognition, the paper presents the outstanding challenges the technique faces for it to match the performance of contemporary linear segmentation methods.

1. What is tiered segmentation?

1.1 The Architectural problem in ASR

Most systems for Automatic Speech Recognition (ASR) that have been constructed since the 1950s have not been based on the principle that speech is some kind of direct implementation of phonology - an equivalent of morse code or DTMF[1] - which simply requires decoding, but rather they have been based on the principle that speech is an impoverished reflection of the activity of the speaker's internal phonological state machine. The nature and sequence of the noises a speaker utters provides evidence for the identity of and transitions between phonological units. This evidence, distorted because of the unique physical properties of a particular vocal tract and acoustic environment; and variable due to the nature of articulation and the linguistic context, does not exhibit the invariant segmented properties of the supposed underlying phonological structure.

Thus modern ASR systems eschew principles of direct decoding - algorithmic functions determined by *a priori* principles which deliver phonological information directly from the signal. Such functions, while providing a parsimonious acoustic-phonetic theory (i.e. they map surface sound to underlying unit with few free parameters), have always been found to be less adequate than functions of a convenient mathematical form with many free parameters which may be estimated by *training*. This observation gives rise to the Architectural problem in ASR:

> *The Architectural problem in ASR is the problem of designing a mathematical framework whereby relationships between signal and message demonstrated by training material may be utilised for the determination of the message of new material.*[2]

[1] Dual-Tone Multi-Frequency, the signalling system used on push-button phones.

[2] For this paper we are concerned with signals and phonological representations only.

From the earliest whole-word systems to modern tri-phone continuous speech systems, researchers have tried to construct models of speech production, estimate the parameters of those models from training data and then determine the most likely inputs to such models given only the evidence of the unknown signal. Modern systems construct the equivalent of gigantic state machines incorporating syntactic, lexical and phonological constraints and use brute-force searching methods to determine the most likely state sequence given an input supposedly generated by an equivalent (human) version of the machine.

The issue that concerns us in this paper, and which is at the heart of the architectural problem, is how to go about relating the underlying phonological structure of a known spoken message to its acoustic form in such a way as to be able to recover the phonological structure of an unknown message.

1.2 Linear phonetics and phonology in ASR

I have argued elsewhere (Huckvale, 1990 & 1992) that acoustic modelling can not ignore the phonological structure of the lexicon. The information that is encoded is intended for lexical access; the choices we need to make are between meaningful linguistic interpretations. Thus we need to do more than (i) model whole words, or (ii) model the noises. Models of words are insufficient because words share functional units; to model words independently is to make multiple models of functionally equivalent units. Misrecognitions can then occur because words will be matched as wholes rather than as an assemblage of units; functionally equivalent units will be in competition contributing irrelevant penalties to the overall score. Models of noises are insufficient because distinctiveness depends on the lexicon and not on the degree of acoustic similarity; large changes in vowel quality may be irrelevant (e.g. 'bath' words), but small changes in frication quality may change the word (e.g. 'three' and 'free'). Thus modelling the acoustics to any particular accuracy will make both too much and too little discrimination.

Contemporary ASR systems utilise acoustic models of linear phonological segments to drive a phonological state machine, a concept dominant since HARPY in 1975 (Lowerre, 1980). In current systems, not only is the phonology linear, but the utterance is recognised as a linear path through a syntax network of word pronunciation graphs. There is only one 'level' of interpretation: that of the word sequence - recovered from the utterance with disregard for the speaker, the accent, the speaker's mood, the rhythm or intonation.

Our particular concern in this paper is at the lower levels: the relationship between sound and phonological description. Conventionally, the sound sequence is modelled as a sequence of 'phones': quasi-phonological linear segments which have variation conditioned on the immediately neighbouring phones. The phones are chosen to at least model lexical choice, but are of a number chosen to represent acoustic variation to a degree of fidelity allowed by the quantity of the training material.

To build linear phone models it is necessary to provide linearly labelled speech signals for training: signals divided into contiguous non-overlapping regions each with a single label. Models of each type of label are then constructed, potentially conditioned by immediately adjacent labels. Once an initial set has been constructed, unlabelled material can be used to further train the models on the assumption that the initial set provides a good-enough transcription alignment.

For recognition, the task grammar and lexical pronunciations are compiled into a directed graph, and for each utterance the single best path through the graph given the phone models and the unknown utterance is determined by an efficient graph-search procedure called dynamic programming.

1.3 Multi-dimensional phonetic representations

The linear phone model of speech makes a fairly direct link between the acoustic properties of the signal and a set of linear phonological units. While a large number of phones is commonly used (many more than are necessary for simple phonemic distinctions), there is only weak attention given to the articulatory structure interposing between phonology and acoustics.

Just as I have argued that acoustic modelling cannot ignore the phonological structure of the message, equally the acoustic modelling should not ignore the articulatory structuring of the signal. It is necessary for the modelling to be sensitive to the acoustic variety of phonological units and this can be best accounted for by considering the processes of articulation.

Despite the continued use of phonetic transcription, it is a view universally held in phonetic science that articulation cannot be viewed as a sequence of discrete gestures, where the transitions between gestures can be safely ignored. Not only are the articulators moving in a continuous smooth motion during speech, but their movement is asynchronous and overlapping. We often use the term *coarticulation* to describe articulatory processes of anticipation and smoothing.

If we seek to describe and model articulation, we need to model articulator movement with a non-linear or multiple-stranded phonetic description: a kind of multi-line graph describing articulator activity, not unlike the 'articulatory descriptions' beloved of elementary practical phonetics teaching. Since ASR needs to be sensitive to the acoustic variety of phonological units this implies that the acoustic signal should be modelled according to their articulatory variety - and hence to the multi-dimensional nature of articulator movement.

Thus contemporary ASR might model the assimilation of lip-rounding or nasality through the use of different phone models for the normal and assimilated forms - still a discrete segmented view of articulation. Whereas to model articulatory variety properly requires a model of how lip-rounding or nasality affects the signal independently of the other articulators. A multi-dimensional phonetic description begs for a multi-dimensional acoustic model.

1.4 Multi-dimensional phonological representations

While the choices that are made in the lexicon to differentiate words can be based on linear phonological units such as phonemes, this is not to say that phonemes play a role in the human decoding process nor that linear segmentation is the only possible way of describing choice. When considering a single uttered syllable such as 'bib', it is just as adequate to talk of the syllable starting with the segment /b/, as it is to say that the syllable nucleus has a bilabial plosive onset. Indeed this latter approach is superior in that (i) we know that the acoustic manifestation of bilabial onsets depends on the following vowel and (ii) the syllable offset, while phonemically /b/, is quite different again.

Similarly, we know that voicing encodes a number of linguistic attributes: not only to differentiate classes of consonants, but also to convey intonation and mood. Our linear-segmented representation is hopeless to convey intonation because voice pitch varies (largely) independently of the segment sequence. In other words voice pitch occupies a different 'strand' of information about the utterance.

Such arguments combined with the need to describe phonological processes in the world's languages in an elegant and parsimonious way have lead to the development of non-linear or 'auto-segmental' phonological models. Such non-linear models replace the segmented linear symbol string with structures in which streams of information about the utterance are changing independently and asynchronously with time. In one view (see Durand, 1990, p257) streams representing intonational tones, vowel tenseness and segmental quality are synchronised via a central consonant-vowel skeleton. Such a structure provides a convenient mechanism for describing the domain of phonological rules which would otherwise involve arbitrary segment re-write rules operating within a heterogeneous set of contexts.

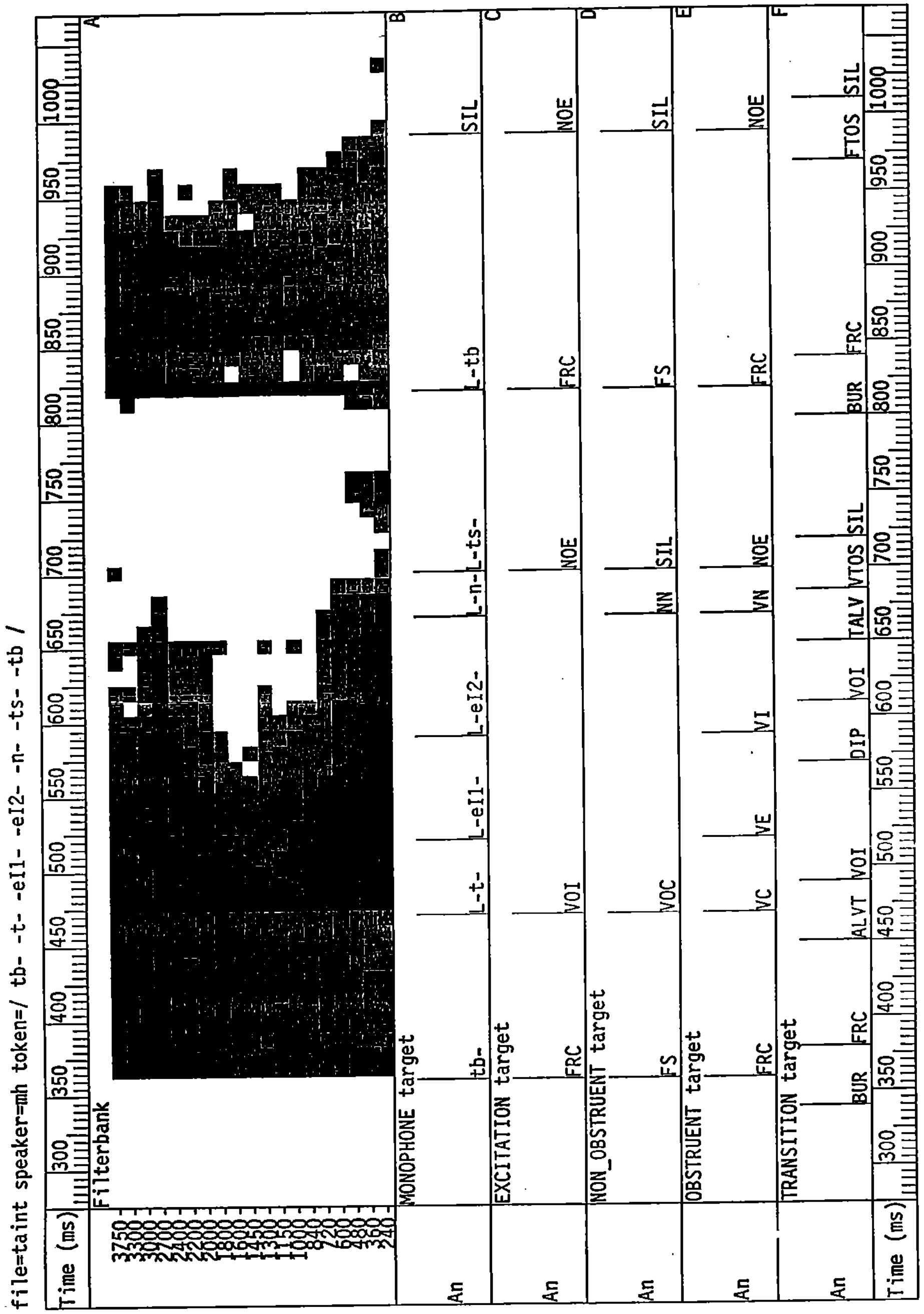
So while in section 1.3 I have argued for multi-dimensional phonetic and acoustic representations in ASR, we may also note that phonologists are themselves investigating alternatives to linear phonological representations for the description of language.

The relationship between the three multi-dimensional representations of speech is a complex matter which is beyond the scope of this paper. My general view of the role phonology should play within a speech recognition architecture is outlined in Huckvale, 1990.

1.5 Tiered Segmentation

A tiered segmentation of a speech signal is then just a multi-dimensional annotation of the signal in which levels of labelling or types of annotation are performed on independent strands or 'tiers'. Figure 1 shows the word 'taint' analysed on levels describing excitation, non-obstruent quality, obstruent quality and transitions. Each tier has an associated inventory of labels or 'elements' and the signal is described as consisting of a non-overlapping sequence of these elements on each tier. Each tier is then independently associated with the acoustic

Figure 1



structure of the signal via some acoustic models of the elements and a grammar of element sequences.

This paper is concerned with acoustic-phonetic mapping and aims to contrast, both theoretically and practically, linear from non-linear segmentation. The discussion in section 2 centres on the potential benefits of a tiered phonetic representation to ASR; section 3 describes the methods by which tiered segmentation might be implemented as a recognition procedure, while section 4 gives a practical demonstration of this. Finally sections 5 & 6 outline the remaining problems and challenges.

2. Why is tiered segmentation useful?

2.1 Representation of acoustic, articulatory and phonological structure

If the object of speech recognition were to recover the position and motion of the articulators from the signal, there is no doubt that we would be building a mathematical vocal tract model .and estimating its parameters (as indeed LPC attempts in a crude way). We would hardly attempt to describe the dynamic vocal tract as a sequence of static articulator configurations and attempt to estimate the sequence of those configurations by looking at their grammar. Such a view would be incompatible with simple observations about the independence and smoothness of articulator movement.

Since however, the object of speech recognition is to retrieve the message from the signal, we find ourselves instead building acoustic models of phonological units. It is necessary for us to distinguish "three" from "free" not because they sound very different, nor because they have different articulations, but because they signal different messages. Simply from a knowledge of sounds or articulations alone we can not build a system that discriminates solely between different meanings.

The architectural problem of ASR is how best to capture the relationship between the signal and the message: conventionally between the signal and its phonological prescription. And although the construction of a system which ignores the articulatory structure of speech seems perverse, this is precisely how systems based on linear segmentation operate.

A tiered segmentation of the speech signal should be an improvement because it recognises the importance of articulation in moulding the acoustic realisation of phonological structures. The consequences of simple articulatory properties such as articulator dynamics should have simple phonological description. If the tongue does not quite meet a target, this should not mean an arbitrary segment substitution. Similarly, small changes in articulation should be reflected in small changes in phonological structure; the weakening of a stop to a fricative for example is a small articulatory step. This sensitivity to articulation in the phonology should lead to better acoustic models of phonological structures since small articulatory steps are linked by physics to small acoustic steps.

2.2 Representation of variability

A second aspect of the sensitivity of tiered segmentation to the articulation of speech is that pattern recognition systems exploit consistency. Speech signals are variable enough - due to linguistic context, environment, speaker or repetition - without that variability being unnecessarily enhanced by inappropriate phonological structures. An incorrect choice will lead to models having large variance and poor discriminative power in recognition. If we start out by saying that syllable-initial /b/ is the 'same' as syllable-final /b/ because our phonology says so, we will end up with a poor acoustic model for /b/ which may well overlap with other inadequately modelled segments.

Thus as well as providing a more sensitive phonological scheme, tiered segmentation also provides a framework in which we can better model articulatory and acoustic variability. One might hope that the effects of linguistic context may be more contained within one tier than spread across all equally; or that some tiers may be more robust to environmental noise, or that speaker characteristics may be better explained by fewer parameters linked to a subset of the tiers, or that repetitions shift only the boundaries between elements in tiers.

2.3 Sharing of acoustic information

In contrast to a linear scheme, in which each segment is modelled independently, the division of phonological information across tiers means that information about acoustic realisation is shared between linear segments. Modelling of the high-frequency fricative spectrum might be shared between /s/ and /z/ for example, because excitation is modelled on a separate tier.

One advantage of this sharing is that there are fewer parameters to train in a tiered segmented recognition system - this means that better models are obtained for a given size of training set.

However the most important aspect of data sharing is not the reduction in parameters but the implicit increase in discriminative power. Separate models of /s/ and /z/ contain two models of the high-frequency fricative spectrum - two models trained independently and with slightly different means and variances. However we know that slight changes in this region of the spectrum do not differentiate /s/ from /z/ and are irrelevant to making the decision between the two phonemes. By explicitly sharing such information in the tiered model, weight is placed instead in the low-frequency region where a voicing decision can be sensibly made.

2.4 Temporal extension of phonetic evidence

In a linear segmentation each section of an utterance is recognised as belonging to one part of one linear segment, even though we know that information about segment identity is spread through the syllable. Formant transitions, for example, provide information both about the consonant and the vowel. While this information is recovered in part through the use of phonotactic constraints, the use of phone-in-context models can only make use of directly adjacent segment influences independently modelled for each phone. A tiered representation allows a region of the signal to contribute to the identity of more than one linear segment. Formant transitions could aid in the identity of a vowel on one tier and to the identity of a consonant on another.

2.5 Annotation without compromise

One final opportunity provided by tiered segmentation is an improvement to the procedures and criteria used for annotating speech data. Currently available linear-segmented speech databases come with documents describing the criteria that the annotators used to divide the signal into contiguous, non-overlapping labelled regions and the compromises that had to be made. Anyone who has experience of annotating signals will know that the choice of symbols and positioning of boundaries are persistent problems (see Barry & Fourcin, 1990).

3. How is tiered segmentation performed?

Given the potential for the description of speech with a tiered segmentation, we may now ask what mathematical procedures are available for modelling and recognising speech on this basis.

3.1 Syntactic Pattern Recognition

Pattern recognition methods can be loosely divided into two classes: those based on the determination of the class (or category) for some unknown set of measurements, and those based on determining the sequence of categories within a sequence of measurements. Traditional whole-word speech recognition falls within the first (decision-theoretic pattern recognition) as it treats words as entities, and uses a complex distance metric (dynamic programming) to accommodate spectral and temporal variability. The introduction of syntax constraints into such recognition fits rather uneasily on top of the word recognition procedure - it adds to rather than is integrated with the metric. The introduction of the Hidden Markov Model framework changed that by using a probability-based distance metric which could be fully integrated with probability-based syntax constraints. This syntactic pattern recognition procedure is inherently more useful for speech recognition where sequence constraints provide essential additional information needed to interpret an utterance.

The issue that confronts us then, is how to make best use of the syntactic constraints of an element sequence within a tier and, importantly, across tiers. This issue does not arise in the linear segmental view, since the constraints that apply are only of a single sequence: and these can be faithfully modelled by conventional syntactic pattern recognition schemes. Let us first consider sequences within a tier.

Compared with the sequence constraints available to linear segments, the constraints are much weaker within a single tier. Given an excitation tier which switches between silence, voice and frication, the six possible transitions between these are approximately equally likely. Also we cannot establish that a given element sequence is impossible as easily as we might establish that a linear phone sequence is impossible. This does not mean that syntactic constraints within a tier are of no use, merely that we need to be mindful in establishing the inventory and grammar for the tier that the greater the sequence constraints the greater the likely recognition performance.

In a previous paper (Huckvale, 1992), I have described the use of two syntactic pattern recognition strategies for tiered recognition: Hidden Markov Models (HMM) and an augmented Multi-Layer Perceptron (MLP) technique. In the first, each tier is described as if it were a broad-phone-class recognition problem, with an inventory of models, a finite-state grammar and a set of transition (bigram) probabilities. In the second, a single MLP is constructed which attempts to categorise 30ms sections of the input signal into one of the element categories. The output of the MLP applied to the whole word to be recognised is then parsed by a dynamic programming procedure to determine the best element sequence that fits the grammar. In this case, bigram probabilities are not used.

3.2 Annotation and Training

For either HMM or MLP it is necessary to provide a quantity of training material which is used to estimate parameters: for the HMM, these are the observation probabilities and the state transition probabilities; for the MLP these are the connection weights. The training material consists of annotated processed speech which demonstrates the association between typically observed speech signals and an *a priori* phonological labelling.

In the conventional linear phone case, the phonological labelling merely identifies the speech signal as consisting of a sequence of non-overlapping contiguous regions each with a single segment name. The inappropriateness of this merely emphasises the ideas behind this paper. Such is the ubiquity of linear methods in ASR that all speech databases available for recognition research are labelled in this way.

For the tiered scheme, we require multiple levels of annotation in which element labels are attached to contiguous, non-overlapping regions within a tier, but where element boundaries do not necessarily align across tiers. We immediately run into problems about how we might obtain a useful quantity of speech labelled in this way since we are multiplying the annotation effort required. In the short term, the only solution is to estimate tiered annotations from linearly annotated material despite the implicit limitations and contradictions in doing so. Fortunately, training methods of HMMs mitigate some of these problems.

To train HMMs on a tiered segmented word, it is satisfactory to train each tier independently. Firstly, all sections of signal annotated with a given element label are extracted and used to train a model of the element (standard HMM parameter estimation). Secondly, each tier in each word is considered in turn and a model sequence is established based only on the element sequence within the tier. This sequence of models is then re-estimated from the entire signal for the word (embedded HMM re-estimation).

The benefit of embedded re-estimation is that at this second stage of training, the only the label sequence is used, rather than exact annotation positions. The problem of embedded estimation is that it becomes necessary to construct HMMs which explain every part of the signal on each tier. Thus while it would have been beneficial to use the vowel quality models to analyse fricatives (e.g., /h/), or fricative models to analyse vowels, the use of embedded estimation - necessary because of the compromised annotation - means that we must use separate models to represent the non-vowel portions during training of the tier. These 'padding' models will always be of poor quality because they cover such a wide range of segments, and they may have a significant effect on the recognition performance within the tier as a whole.

For the MLP technique, training only consists of identifying 30ms sections of signal as belonging to one of the element classes within the tier. In contrast to the HMM technique, the MLP constructs a procedure for discriminating between elements, rather than a procedure for choosing the most likely. The MLP is a single model of all the elements in the tier; it has one output for each element class. During training it must modify its internal structure to differentiate between elements.

Thus the benefit of the MLP is in its discriminative power and its ability to label 30ms sections of the signal with the correct label. The problem of the MLP is that its outputs do not directly correspond to probabilities, so that it is difficult to establish from the table of outputs for an entire input word which element sequence is most likely given also the grammar for the tier and element transition probabilities.

We can test tiered recognition methods on two levels: whether the elements and element sequences within the tiers match the annotated reference and whether the system recognises the utterance as a whole. The methods for syntactic pattern recognition described above only provide independent estimates for the contents of each tier. Thus we can certainly recognise and compare performance within tiers. However we have not yet described procedures for utterance recognition by combining information across tiers. We describe the problems of doing so in Section 5.2.

3.3 Performance Measures

Tiered segmentation also provides challenges for measures of recognition performance. Because, as yet, there are problems of building effective sentence or word recognition systems based on tiered segmentation, measures such as 'phoneme' accuracy have to be adapted to compare conventional linear segmentation recognition performance with tiered segmentation.

The conventional measure of linear performance is 'accuracy' which is calculated from two label sequences by subtracting the percentage insertion rate from the percentage correct rate. To do this requires that the two label sequences are aligned to minimise the number of insertions, deletions and substitutions required to transform the recognised sequence to the correct sequence. If the number of labels in the correct sequence is N, the number of correctly recognised labels (after alignment) is C, and the number of recognised labels that have no reference counterpart is I, then the accuracy is 100*(C-I)/N.

We can use accuracy within a tier to give a performance measure for that tier alone, but we are immediately faced with the possibility that a recogniser could have extremely high performance on each tier, but still provide a poor recognition of the phonological structure because the alignment between tiers may be highly variable from one repetition of a word to the next. We require our recognition system to provide a suitable phonetic description overall not on each tier independently.

An alternative measure is to compare the recognised labels with the reference annotations not only in terms of element name but also in terms of element position. A crude measure is 'frame-labelling' performance, which indicates the percentage of 10ms signal frames which are labelled correctly. This provides a measure which is sensitive to label alignment, but at the penalty of giving undue emphasis to long segments.

Interestingly, the HMM system is better designed to provide a recognised label sequence than a set of frame labels, while for the MLP the situation is reversed.

When comparing tiered recognition against linear methods, we can always 'map' down linear results into tiers so that comparisons may be made. The inherent danger of this is that the reference annotations have also been mapped down - so that the linear system has an in-built advantage in terms of cross-tier alignments.

A third performance measure can be found by determining, for each tier, how many utterances were labelled correctly; i.e. with the correct sequence of element labels. We can predict that a system which obtains high 'phrase labelling' performance for each tier will have a high phrase recognition performance for the utterances, since utterances always differ by the element sequences taken over all tiers.

Finally, when comparing different tiered recognition methods any performance comparisons within tiers may be useful, but for comparison with conventional linear recognition methods only word recognition performance is going to be an acceptable measure.

4. Does tiered segmentation work?

From this rather abstract discussion of the nature of tiered segmentation and methods for its implementation, we now turn to a concrete example of tiered recognition. This is a first implementation of the ideas, which has been useful in demonstrating the outstanding challenges facing tiered segmentation. A more complete description may be found in Huckvale (1992).

4.1 Example database

The MONOS database is designed to explore the phonetic variety of isolated monosyllabic English words. The particular vocabulary used in this experiment looks at a large subset of permissable English initial consonant clusters, a large subset of permissable English final consonant clusters and a large subset of permissable nuclear vowels.

The 46 initial consonant clusters chosen were[3]:

NULL, b, d, g, p, t, k, m, n, l, r, w, j, dZ, tS, f, s, S, T, v, z, D, h, bl, br, dr, gl, gr, pl, pr, tr, tw, kl, kr, kw, fr, fl, sp, st, sk, sl, sm, sn, sw, Sr, Tr.

The 15 vowels chosen were:

i:, I, e, {, V, A:, O:, Q, u:, 3:, aI, eI, OI, @U, aU.

The 48 final consonant clusters chosen were:

NULL, b, d, g, p, t, k, m, n, N, l, tS, dZ, f, s, S, T, v, z, bz, dz, gz, ps, ts, ks, mz, mp, nz, ns, nt, nT, ntS, ndZ, Nz, Nk, lf, lz, lp, lt, lk, fs, sp, st, sk, vz, ft, ld, ns.

667 English words were then found which exercised most legal possibilities of each initial cluster followed by each vowel, and separately each final consonant cluster preceded by each vowel. This became the training set. A further 359 English words (not present in the training set) were then found which re-covered approximately 50% of the consonant-cluster/vowel combinations in the training set. This became the test set. The word lists are available from the author.

Recordings were made of a single male speaker with a close-talking microphone in an office environment. Automatic end-pointing based on energy criteria was used to isolate each word; items that were too quiet (used fewer than 11-bits of the ADC) or overloaded (used more than 12-bits of the ADC) were automatically rejected. Each recorded signal was also quickly inspected at the time of recording and a minority of utterances (less than 10%) were rejected and re-recorded.

[3] Transcriptions are printed in SAMPA notation.

The signals were annotated using an inventory of 117 sub-phonemic labels. The inventory was chosen to (i) identify important acoustic changes in the signal, (ii) label phonological distinctions, (iii) separate potential contextual variants of phonological units. So stops were divided into burst, gap and vowel-transition regions; fricatives /T-/ and /f-/ were given separate labels; /r/ was labelled differently after /t/ than as a separate syllable onset. The annotation of the words was performed by an automatic dynamic-programming (DP) alignment between the signal and a concatenated sequence of spectra for each annotation label taken from a hand-generated dictionary.

We report on results obtained from a vocoder-based feature vector comprising 19 filterbank energies relative to the overall energy and the overall energy value itself (based on the filters used in the JSRU vocoder, Holmes 1980). The frame rate was 100s^{-1}. The recordings and annotations are available from the author.

4.2 Example tiers

The EXCITATION tier divided the words into three classes of region:

NOE	No excitation
FRC	Mainly fricated (aperiodic) excitation
VOI	Mainly voiced (periodic) excitation

Voiced fricative regions were allocated to the FRC class.

The NON-OBSTRUENT tier attempts to label the primarily voiced, non-obstruents: those that could be expected to show a clear steady-state formant structure.

NOE	No excitation	FRC	Mainly fricated excitation
VI	Front, close vowels and /j/	VE	Front, half-open vowels
VH	Front, open vowels	VU	Back, close vowels and /w/
VO	Back, half-open vowels	VA	Back, open vowels
VR	Central vowels and /r/	VL	Alveolar lateral
VN	Nasals	VC	Voiced obstruents

The OBSTRUENT tier attempts to differentiate obstruents, primarily fricatives, bursts and nasals. The elements are:

SIL	Silence	VOC	Non-obstruent voicing
FP	Bilabial frication	FF	Labial frication
FS	Alveolar frication	FSH	Palatal frication
FX	Velar frication	FH	Glottal frication
NM	Labial nasal	NN	Alveolar nasal
NX	Velar nasal		

The TRANSITIONS tier attempts to differentiate between different types of spectral transition in the signal:

SIL	Silence	STOF	Silence to Frication
STOV	Silence to Voicing	FRC	Frication
FTOS	Frication to Silence	FTOV	Frication to Voicing
LABT	Labial opening	ALVT	Alveolar opening
VELT	Velar opening	TLAB	Labial closing
TALV	Alveolar closing	TVEL	Velar closing
BUR	Stop burst	APP	Approximant
DIP	Diphthong		

To generate annotated regions for these transition labels, the 117 monophone labels were first mapped to a set of broad classes and then 40ms transition regions were labelled at each broad class junction. All resulting regions were then mapped to one of the classes above.

4.3 Method

Hidden Markov Models were trained and tested using the Cambridge HTK software vs. 1.2 developed by Steve Young. All element models had three emitting states with single Gaussian mixtures, diagonal covariance and self+next transitions only. The models in each tier were first independently initialised and re-estimated and then fine-tuned with 5 cycles of embedded re-estimation.

For recognition, each tier was allocated a syntactic network which specified legal sequences of elements within each isolated word. The design of the network was not based on the specific training and testing vocabulary, but on the broader design goals of the MONOS database subset, that of 46 initial consonant-clusters x 15 vowels x 48 final consonant clusters. The only compromise to this very general position was to prevent short vowels occurring in open syllables. Bigram probabilities were collected for each tier from the entire 1026 (667+359) word vocabulary.

Multi-Layer Perceptron models were trained and tested using the Pattern Recognition Workbench (PRW) tools developed at UCL. All models take three adjacent input vectors (3x20 values), have a single hidden layer and an output layer of a size determined by the number of elements in the tier. The number of units in the hidden layer was chosen to be twice the number of units in the output layer. By this means, the total number of weights in the model approximated the total number of parameters in the parallel collection of Markov models for the tier. For each input vector triplet the training vector consisted of a value of 0.9 for the labelled element output and 0.1 for the others.

The models were trained using an adaptive back-propagation technique with weight updates every 50 vectors presented. Models were trained for 20 complete passes over the training data, by which time the residual squared error change per cycle was always very small.

The models were used for recognition by first performing a forward pass over each isolated test word to generate a vector of output values for each input frame. A DP procedure then generated a legal element sequence for the tier, constrained by a simple syntactic network as used for the HMM models. The distance measure chosen between an element **e** on the network and the MLP output **o(e,t)** at time **t** was simply:

$$d(e,t)=\frac{\sum o(i,t),i\neq e}{\sum o(i,t)}$$

4.4 Results

We start by establishing a base system, comprising HMMs of linear phones, one model for each annotation label used for the vocabulary. These have the same structure as the element models and are trained in a similar fashion. For recognition, we use a syntax based on the design of the MONOS database, allowing any initial consonant cluster to precede any vowel, and any final consonant cluster to follow any vowel. By mapping down the recognised sequences into tier labels, we are able to obtain performance measures which are comparable to the tiered recognition procedure.

We now present, for each tier in turn, three performance measures: (i) element label accuracy, (ii) frame labelling percentage correct, (iii) phrase label percentage correct. We show baseline results, then results for the HMM and the MLP systems.

EXCITATION	Label Accuracy%	Frames Correct%	Phrases Correct%
Baseline	96.5	91.9	85.3
HMM	86.1	90.1	50.4
MLP	88.8	92.6	52.7

NON-OBSTRUENT	Label Accuracy%	Frames Correct%	Phrases Correct%
Baseline	92.8	87.8	63.8
HMM	81.3	84.6	26.5
MLP	75.6	86.6	15.3

OBSTRUENT	Label Accuracy%	Frames Correct%	Phrases Correct%
Baseline	91.9	89.8	64.9
HMM	79.8	86.1	29.3
MLP	75.0	85.6	16.4

TRANSITION	Label Accuracy%	Frames Correct%	Phrases Correct%
Baseline	89.0	83.2	39.6
HMM	66.3	68.6	3.3
MLP	72.9	82.6	5.3

Thus for all but one of the tiers, the best performance regardless of performance measure is obtained by linear phone labelling followed by mapping of those phone labels into element sequences.

4.5 Discussion

There are a number of points that should be made about these results:

1. The performance of the baseline system is strongly dependent on the use of the bigram probabilities obtained from the train and test vocabulary. To emphasise this the EXCITATION tier results without bigrams looks like:

EXCITATION	Label Accuracy%	Frames Correct%	Phrases Correct%
Baseline (no bigram)	58.9	86.6	1.1
HMM (no bigram)	61.9	87.0	4.2
MLP	88.8	92.6	52.7

Without the use of the bigram probabilities, the baseline error rate degrades 12-fold. The HMM tiered recognition degrades 3-fold, and the MLP system (which does not use bigram data) is unaffected.

2. As mentioned in 3.3, the 'mapping down' of the baseline linear segmentation results into tiers gives an inbuilt performance advantage when the tier labels have themselves been generated from linear annotations. Both the reference tier labelling and the mapped down linear labelling are

constrained to have boundaries that align across tiers. An annotation scheme which did not force alignment across tiers would, in contrast, put the linear system at a disadvantage.

3. Apart from the excitation tier, all the tiered HMM recognition models use 'padding' models: HMMs which are used to fill-in regions of the word for which the tier offers no explanation (the VC element in the non-obstruent tier for example). In fact we would prefer not to include this model at all in recognition; but as yet we have no mechanism for performing recognition without it (what would the grammar for the tier look like?) and no performance measure that is set up to ignore the padding labels.

5. Can tiered segmentation deliver?

From this diversion into the practicalities of performing tiered segmentation, we can now describe the challenges ahead.

5.1 The design of tiers

We can see that the design of the tier structure is influenced by many variables: (i) the phonological model, (ii) the ability to annotate, (iii) the need for strong sequential constraints (both grammatically and probabilistically), (iv) the avoidance of padding elements.

The tier structure in section 4 has a number of weaknesses: it divides the tiers unnecessarily between vowels and consonants; information is repeated across tiers (all vowel elements in the non-obstruent tier have a corresponding VOI element in the excitation tier); transitional information is banished to an independent tier. The removal of the duplication in the models across tiers should have two benefits: improved discrimination and more tier independence; both leading to a better model of phonetic variability within phonological constraints.

A linear annotation which must be mapped down to give tier labels provides an awkward basis from which to train tier models. Each transitional element in the transition tier in Section 4, for example, was forced to be 40ms long so that it could be predicted automatically from the boundary between linear annotation labels. Not only is this a poor approximation to labelling, it also gives the linear results a huge advantage when the recognised label boundaries are also mapped down to 40ms duration. The choice of elements must also then be influenced by what annotation scheme is available for the training material.

The most important outcome of the experimental results from Section 4 is the demonstration of the influence of sequential constraints in improving recognition performance. The baseline linear system, having a larger inventory of recognition units exploits constraints on unit sequences provided by the 1000 word vocabulary much more effectively than can the tiered system with a small inventory of element units within each tier. To make more effective use of sequential constraints in the tiered system, we need to choose a larger inventory with non equal transitional probabilities. However, it is also worth noting that as the existing system moves to a larger vocabulary or to connected utterances, the sequential constraints will weaken for the linear system. The hope is therefore, that the combination of an improved tier inventory and a more difficult recognition task will lead to a convergence in performance.

Finally the use of padding elements in the tiers is to be avoided. This might be done by simply using a different inventory of symbols for training as against recognition. The padding models (models of a variety of segments not explicitly described by the tier) would only be necessary to allow embedded re-estimation of HMM tier models; but would then be discarded. At recognition time the grammar and transition probabilities would need to take into account the 'true' inventory for the tier, and the performance measures adjusted to only consider the relevant sections of the recognised label sequence. Since the foregoing seems rather complex, it may still be better to design the tier structure to give complete coverage within each tier - this may be required anyway to give good sequential constraints.

5.2 Word matching strategies

The second major challenge for tiered segmentation is the step to word recognition.

The simplest approach is to view the phonetic recognition and the lexical choice as independent steps. Thus in the experiment in section 4, we could take either the linear transcription from the baseline system, or the transcriptions from the four tiers to match against phonologically-predicted entries for the words in the test vocabulary. This matching is based on a metric which returns the distance between the recognised transcription and the predicted transcription, so that the closest word may be chosen.

We have implemented an extremely simple version of this idea, by using a dynamic programming metric which matches each tier transcription in turn to a vocabulary word, with penalties for the insertion and deletion of elements and for temporal distortion, and sums the distances across tiers. This gives a truly abysmal word recognition performance. For comparison the table below also includes the mapped down baseline results used for recognition with the same procedure:

System	Words Correct%
Baseline (mapped)	61
HMM	37
MLP	26

A major problem is that the division between phonetic recognition and lexical choice ignores the very strong constraints provided by the limited vocabulary that could and should be applied-at the lowest level. Thus, we need a single recognition procedure that delivers the word identity at the same time as delivering a transcription. For word recognition, we must use information about the specific phonological structure of the test vocabulary in the phonetic recognition process.

This may be readily demonstrated for the baseline system by substituting the recognition grammar (that normally allows any combination of initial consonant cluster, nuclear vowel and final consonant cluster) with a grammar composed only of the transcriptions of the test vocabulary. Since each legal transcription is now a test word, we can recover word recognition results directly. The baseline system constrained in this way delivers a word recognition rate of 96%!

Thus what is required is to construct a similar scheme for the tiered recognition case. However, the multi-dimensional nature of the transcription makes this tricky to achieve. Whereas the parsing of a linear transcription involves finding the best path through a directed graph, the parsing of an N-dimensional transcription involves finding N paths through N graphs (where timing correspondences are forced across graphs) such that the combination of paths is 'best'. A solution to this is still awaited.

Another issue of word recognition, largely ignored in the linear case, but important in the tiered case is the modelling of phonetic variation within a word. In the linear case, phonetic variation has been modelled by constructing alternative transcriptions for words and including these in the vocabulary: thus "and" may have transcriptions /and/ and /an/, or even /n/. Notice that pronunciation variety is limited to phonological unit variation. In the tiered case, we expect the alignment between elements across the tiers to vary as articulation varies: we also expect some elements to be substituted or deleted according to contextual, speaker or environmental variation. This type of variation needs to be modelled at a finer level of detail than in the linear case: with analysis needed of the typical durations and boundary variability for each element in a number of contexts. We would expect that this knowledge would also contribute to the distance metric in a multi-dimensional parsing for word recognition.

6. What is the future for tiered segmentation?

In this paper we have tried to show the challenges facing tiered segmentation as well as the opportunities. In Section 2, we have described the potential for modelling the phonetic structure of speech signals with a multi-dimensional phonetic transcription. In Section 3, we have described some of the methods available to perform speech recognition based on a tiered segmentation, while Section 4 demonstrates some of these ideas in practice. In Section 5, we have outlined the major problems facing the technique. Finally, we shall consider where effort needs to be concentrated.

In terms of choosing the inventory of tiers and elements, it is imperative to exploit sequential constraints within tiers to maximise the performance of syntactic pattern recognition procedures. In addition it will be beneficial to ensure that the elements in each tier cover the entire signal in a useful way - that no elements are just padding. It is believed that tiers that distribute information uniquely across tiers will have better performance, so that duplication of information across tiers (dependency) is best avoided.

In terms of lexical access, it is essential to develop parsing methods that combine tier segmentation with lexical access: this means that vocabulary limitations can be used to constrain phonetic transcription. Bigram and other statistical measures have been shown to be useful for the HMM systems and should be incorporated into the MLP based system.

Methods of annotation also need to be re-considered, as the linear scheme of annotation works against tiered segmentation and actively weakens tiered recognition performance in comparison with linear recognition methods.

7. References

Barry, W.J. & Fourcin, A.J. (1990) Levels of Labelling. *Speech, Hearing and Language, Work in Progress*, Phonetics and Linguistics, University College London, vol. 4, 29-44.

Durand, J. (1990). *Generative and Non-Linear Phonology*, Longman.

Holmes, J.N. (1980). The JSRU 19-channel vocoder. *IEE Proc.*, 127 part F, No.1

Huckvale, M.A. (1990). Exploiting speech knowledge in neural nets for recognition, *Speech Communication*, p1.

Huckvale, M. (1992) A comparison of neural-network and hidden-Markov-model approaches to the tiered segmentation of continuous speech. *Proceedings of the Institute of Acoustics, Speech and Hearing Autumn Conference*, vol. 14, part 6, 165-172.

Lowerre, B, Reddy, R. (1980). The Harpy speech understanding system. *Trends in Speech Recognition*, ed. W.Lea, Prentice Hall.

THE IDENTIFICATION OF /l/ AND /r/ BY JAPANESE EARLY, INTERMEDIATE AND LATE BILINGUALS

Junko NAKAUCHI

SPEECH, HEARING AND LANGUAGE: WORK IN PROGRESS
U.C.L. No 7 (1993)