Proceedings of the Institute of Acoustics

A COMPARISON OF NEURAL-NETWORK AND HIDDEN-MARKOV-MODEL APPROACHES TO THE TIERED SEGMENTATION OF SPEECH

Mark Huckvale

Dept. Phonetics and Linguistics, University College London, Gower Street, London, WC1E 6BT

1. INTRODUCTION

The use of acoustic models of primitive phonological entities called 'phones' is predominant in contemporary Automatic Speech Recognition (ASR) systems. Every unknown utterance is assumed to be describable in terms of a linear sequence of non-overlapping phone models constrained by pronunciation graphs derived from the task lexicon and task syntax. The phonological justification is based on a taxonomic phonemic analysis for the language as a whole (since if an analysis is sufficient to describe contrasts in the complete lexicon then it is at least sufficient for the task lexicon), although such an analysis is performed without objective measures of the acoustic similarity of the phonological segments (e.g. syllable-initial /b/ lumped with syllable-final /b/). Because of the huge acoustic variety of phonemic units (depending on speaker, context and repetition) practical systems divide each phoneme into a set of context-dependent phones, and train separate phone models in different contexts.

There are two important benefits to this:

1. The phone-in-context models have smaller variance without giving rise to additional competing recognition units (since there can be no phone competition within a phoneme class).
2. The phone models are tied to larger contexts than phonemes and can so exploit some of the phonotactic constraints arising from the task lexicon and syntax (providing the task lexicon is small and the test language model is the same as the training language model) [1].

There are still a number of weaknesses however:

1. Models of articulatorily-related segments are modelled independently; this forces up the amount of training data required and leads to poor discrimination.
2. Annotation of training material with linear segmentation is often difficult and of variable quality.
3. Minor changes in articulator timing have large effects on the linear transcription (e.g. start/stop of laryngeal vibration, asynchrony of velum, tongue-tip and lips).
4. The realised form of an utterance shows many context-sensitive changes over the lexically-derived transcription.

What appears to be the root of these weaknesses is the initial choice of a linear phonological basis to the modelling. We are imposing a phonological model of lexical contrast on the acoustic data rather than using an articulatory phonetic model to structure the acoustic data and leaving lexical choice up to the lexicon [2].

2. PHONETIC TIERS

Our previous work (e.g. [3]) has been concerned with generating continuously-valued multi-dimensional feature representations of signals for use as an alternative 'front-end' to conventional recognition algorithms. These have been found to be useful in 'normalising' both acoustic environment and

speaker characteristics. In comparison with current phone modelling systems, these representations operate at an intermediate level between acoustic frames and phones: their output is phonetic but not segmented. As a consequence, lexical access requires a separate syntactic pattern recognition procedure.

This paper explores another possibility: a multi-dimensional segmented representation; one where linear segmentation is used in a number of parallel 'horizontal' layers called *tiers*. Since the tiers divide the description of the signal in an articulatorily-principled way, the segmented tier allows 1. the sharing of acoustic information over phonological segments, 2. less compromise in annotation, and 3. asynchrony and overlap of articulator effect. A tiered segmentation might also help in undoing the convoluted context sensitivity of phone sequences found in connected utterances (see related ideas in Government Phonology [4]).

The general issues about tiered segmentation that need to be addressed are:

1. Does the segmentation into different tiers provide more accurate phonetic information about the signal than a single linear phone segmentation?
2. Does the tiered segmentation provide a more reliable structure for representing the phonetic variability of words and hence improve lexical access (word-recognition) performance?

In this paper we look at a simple implementation of the tiered segmentation approach to the phonetic interpretation of the speech signal. Using just four tiers: Excitation, Non-Obstruent, Obstruent and Transition, each tier is independently recognised using a small inventory of 'elements' within each tier. For example within the Excitation tier, the elements are (roughly): silence, frication, voicing; and the excitation tier consists of contiguous, non-overlapping regions labelled with one of these three elements.

Thus this paper addresses only the first of the two issues, although of course the second also needs to be answered before the whole technique can considered to be useful.

3. METHODS AND DATA

3.1 Database Design

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[1]:

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.

[1]Transcriptions are printed in SAMPA notation.

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.

3.2 Recordings and Annotation

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.

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. The recordings and annotations are available from the author.

Two representations were chosen for the pattern recognition: **Cepstral** is a 13-element vector comprising 12-parameter cepstral coefficients and overall energy; **Vocoder** is a 20-element vector comprising 19 filterbank energies relative to the overall energy and the overall energy value itself. The first has been used extensively in phone modelling experiments, the second used extensively in isolated word recognition experiments (based on the filters used in the JSRU vocoder [5]). Both had 10ms frames.

3.3 HMM Tools

The 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 syntax 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.

Where bigram (model-pair) sequence probabilities are stated in the text, these refer to the collection of bigram data from the specific 1026 (667+359) word vocabulary used in the experiment. Recognition results are stated below with and without the use of bigram probabilities.

3.3 MLP Tools

The 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 (3x13 or 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 two times 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 syntax 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. EXPERIMENTAL RESULTS

4.1 Monophone segmentation

To act as a baseline for the judgement of labelling performance within each tier, HMM-based models of each of the 117 annotation labels were constructed and tested. Only a monophone scheme was used, but the annotation set contains some of the context-dependence of typical bi-phone modelling.

The usual method of measuring phonetic transcription performance is to align the recognised label sequence with the correct label sequence using a DP scheme and report the number of substitutions, insertions and deletions. Doing this for the monophone baseline system using only the general consonant-cluster/vowel/consonant-cluster syntax network gave test performance of:

Cepstral:	Correct=57.3%, Insertions=50.2%, Accuracy=7.1% (N=2398)
Vocoder:	Correct=58.0%, Insertions=49.7%, Accuracy=8.3% (N=2398)

The weakness here was the generality of the network, with weak spectral regions at the start and end of voicing being labelled with inserted fricative regions. The use of bigram probabilities for the

experimental vocabulary confirms this by reducing the number of insertions significantly:

Cepstral+Bigram:	Correct=82.2%, Insertions=3.5%, Accuracy=78.7% (N=2398)
Vocoder+Bigram:	Correct=85.2%, Insertions=3.6%, Accuracy=81.6% (N=2398)

This type of measurement of transcription accuracy is not really relevant in the tiered segmentation case, since the alignment between elements in different tiers is also important. In other words it is not sufficient merely to find out that a word comprises a voiced region surrounded by two fricated regions, because the timing of the change between frication and voicing may affect the phonetic interpretation of the word when taken in combination with descriptions of fricative quality in a different tier. So an alternative way of judging performance, more appropriate for tiered recognition, is the frame-labelling performance: that is the measurement of the substitution error rate only for each input frame label. This measure has to be treated with care since it is now biased by segment duration, and by silence in particular. However we can use frame performance to compare linear segmentation with tiered segmentation for a given data set.

The frame labelling performance for the monophone models on the test data was (N=50941):

Cepstral:	Correct=54.6%
Vocoder:	Correct=59.0%
Cepstral+Bigram:	Correct=78.2%
Vocoder+Bigram:	Correct=82.5%

These are close to the % correct figure for the DP alignment performance. Using these monophone frame-labelled regions we can generate a minimum level of performance necessary to show improved extraction of phonetic data. For tiers in which there is a simple relation between the monophone label and an element in the tier, we can map the recognised monophone labels down to element labels and measure frame performance against the target element labels. The resulting mapped performance figures are reported along with performance using specific tier models in the sections below.

4.2 The Excitation tier

The most primitive 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 Excitation tier results were as follows:

EXCITATION (N=50941)	Cepstral (%Correct)	Vocoder (%Correct)
HMM	89.5	87.0
HMM + Bigram	90.7	89.2
MLP	88.1	92.3
Monophone	82.1	86.6
Monophone + Bigram	90.5	91.9

HMM and MLP are the frame labelling performance for the HMM and MLP models using only the general word syntax. Monophone is the mapped results from the linear phone recognition. Here, as we shall see elsewhere, the use of the bigram probabilities with the monophone models make a significant increase in their frame labelling performance; while the use of bigram probabilities within the tier makes only a small improvement.

4.3 The Non-Obstruent Tier

This 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 recognition results are as follows:

NON-OBSTRUENT (N=50941)	Cepstral (%Correct)	Vocoder (%Correct)
HMM	82.0	81.3
HMM + Bigram	84.2	84.0
MLP	81.0	86.0
Monophone	70.2	76.0
Monophone + Bigram	84.8	87.8

4.4 The Obstruent tier

This 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 results are as follows:

OBSTRUENT (N=50941)	Cepstral (%Correct)	Vocoder (%Correct)
HMM	82.2	81.2
HMM + Bigram	86.5	85.7
MLP	84.5	85.5
Monophone	79.3	83.1
Monophone + Bigram	88.2	89.8

4.5 The Transitions tier

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

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

To generate annotated regions for these 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 in turn to one of the classes above.

TRANSITION (N=50941)	Cepstral (%Correct)	Vocoder (%Correct)
HMM	38.5	55.8
HMM + Bigram	45.3	66.9
MLP	80.4	82.6
Monophone	67.0	72.3
Monophone + Bigram	81.4	83.2

5. DISCUSSION

Considering first the recognition scores found without use of bigram statistics: i.e. recognition with general syntactic constraints within each tier but not specifically trained to the particular training and testing vocabulary. In all tiers the MLP tiered segmentation extracts more accurate phonetic information than the simple monophone system mapped down to the same labels. This was also true for 5 out of the 8 HMM-based tiered segmentation results, however except for the obstruent tier, the best performance was always from the MLP technique.

When the recognition scores using the bigram probabilities are considered: i.e. when we consider the recognition of the specific vocabulary, the results are more mixed. Since the bigram probabilities are more powerful in constraining phone-sequences than they are in controlling element sequences within a tier, the monophone performance was more similar to the best tiered segmentation result. Only in the Excitation tier did the MLP system outperform the best Monophone + Bigram system. The performance within each tier using bigram probabilities could be improved by minor changes to the inventory of elements within each tier (based on cross-confusion matrices with annotation labels), the generation of more specific element-sequence constraints and temporal constraints for use with MLP recognition and segmentation.

For word recognition, the tier-segmented data needs to matched to statistical models of pronunciation variation generated from the standard lexical form of words. Simple Markov Model techniques would be appropriate but we are keen to explore variants of a connectionist model of lexical access [2].

6. REFERENCES

[1] D B PAUL, J K BAKER & J M BAKER, 'On the interaction between the true source, training and testing language model, *IEEE Proc. ICASSP-91, Toronto*, p569-72.

[2] M A HUCKVALE, 'Exploiting Speech Knowledge in Neural Networks for Recognition', *Speech Communication* 9 1990, p1-13.

[3] M A HUCKVALE & J S HOWARD, 'Phonetic Feature Analysis for a Monosyllabic Recognition task', *Proc. IOA Conf. Speech and Hearing, Windermere*, 1990

[4] J D KAYE, *Phonology: A Cognitive View*, Lawrence Erlbaum Associates, 1989.

[5] J N HOLMES, 'The JSRU Vocoder', *IEE Proceedings*, 127 Part F, No.1 1980.