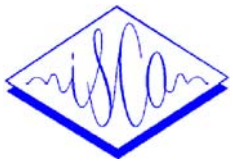

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

CONTINUOUS SPEECH RECOGNITION USING SYLLABLES

Rhys James Jones[1], *Simon Downey*[2], *John S. Mason*[3]

[1,3] Speech Research Group, Department of Electrical & Electronic Engineering,
University of Wales Swansea SA2 8PP, UK

[2] Speech Technology Unit, BT Laboratories, Martlesham Heath, Ipswich, Suffolk IP5 7RE, UK

R.J.Jones@swansea.ac.uk, downey@saltfarm.bt.co.uk, J.S.D.Mason@swansea.ac.uk

ABSTRACT

The vast majority of work in continuous speech recognition uses phoneme-like units as the basic recognition component. The work presented here investigates the practicability of syllable-like units as the building blocks for recognition. A phonetically annotated telephony database is analysed at the syllable level, and a set of syllable-based HMMs are built. Refinements including the introduction of syllable-level bigram probabilities, word- and syllable-level insertion penalties, and the investigation of different model topologies are found to improve recogniser performance. It is found that the syllable-based recogniser gives recognition accuracies of over 60%, which compares with 35% as the baseline accuracy for monophone recognition. It is envisaged that practical applications of syllable recognition could be in a hybrid system, where the most common syllable HMMs would be used in conjunction with whole-word and phoneme models.

1. INTRODUCTION

Continuous word recognition using grammars comprising phoneme-like units poses many practical challenges. The boundaries between phoneme-like units are often difficult to elicit, which can give rise to a lack of context-dependent effects in models of short duration phonemes. Context sensitive and whole-word models can help to circumvent these problems, but imply greater amounts of training data and computation.

In an attempt to avoid the limitations of phoneme-like models without significantly increasing computational overhead, syllable-based HMMs are used to build whole-word grammars. Continuous syllable recognition is also examined.

To date, there has been only a small amount of published work on recognition using syllable units. Hu et al. utilise syllable-like units in an English recogniser having a twelve-word vocabulary comprising the months of the year [1]. Our approach differs from theirs in that a far larger vocabulary is used, and a trajectory feature estimation stage is not included. Recently, Bourlard and Dupont [2] mention the use of syllables in their multi-path experiments, using parallel models, in the German language.

An attempt at defining a syllable, and some experimental background, is found in Section 2. Results are presented in Section 3, together with comparisons with experiments using phoneme-like grammars, a discussion of which is found in Section 4. Section 5 discusses areas for further experimentation.

2. EXPERIMENTAL BACKGROUND

2.1. Syllable definition

Syllables are perhaps easier to identify than to define. It is common to use peaks of sonority or peaks in prominence within an utterance in attempting such definitions [3]. However, a precise phonetic definition of a syllable, as presented by Laver [4], is illustrated in Figure 1. This shows a syllable comprising a central vowel, or vowel-like consonant, called the **nucleus**. The nucleus may be optionally prefixed and suffixed by one or more consonants, termed the **onset** and **coda** respectively.

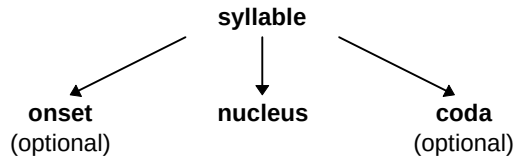

Figure 1: Illustration of syllable components

Syllable boundaries can be determined using the **maximum onset rule.** This states that the intra-word syllable boundaries are placed so as to maximise the number of consonants at the beginning of each syllable (i.e. in the onset). For instance, when applied to the word 'banana', the maximum onset rule states that the syllable divisions are /ba·na·na/, rather than, for example, /ban·an·a/.

2.2. Database

The database used for the work is *BT*'s read speech database *Subscriber* [5], which contains recordings of 1250 speakers, each reading 5 out of 200 different phonetically rich sentences. *Subscriber* contains a vocabulary of 1313 different words. In the work presented here, the training set comprises 3063 sentences, and there are 1810 sentences in the testing set. There are 750 speakers in the training set, and 500 different speakers in the testing set. Recognition is thus speaker-independent, with no adaptation.

Subscriber is hand-annotated with time-aligned phoneme-level transcriptions of each utterance, using the 74 most common British English phonemes in SAM Phonetic Alphabet [6] notation. Using a syllable-delimited dictionary, syllable-level time-aligned transcriptions are extracted from these using forced recognition. In forced recognition, a syllable grammar is prepared for each of the syllables in the database - the grammar for each syllable contains its constituent phonemes. It is not possible to directly utilise the phoneme-level transcriptions, as these are based on the caller's actual utterance, rather than the

dictionary transcription of the sentence. Hence there is no direct relationship between the information available at the phoneme and syllable levels.

2.3. Experiment design

Performance is assessed for both syllable and whole-word recognition. Experiments are also designed to investigate the variation of recognition performance with the number of mixtures in the HMM, and different HMM topologies.

Prototype HMMs are built for each syllable in *Subscriber*. The number of states in each HMM is proportional to the number of phonemes in the syllable. Initial experiments are carried out using three mixture models with no skips. A second set uses one mixture, no skip models.

The amount of training examples for each syllable varies from 9 to 3290, meaning that not all HMMs can be trained to the same level. In order to better reflect the range of training examples available for each syllable, 'stepped mixture' models are also examined. The composition of these varies according to the number of training examples for the modelled syllable in *Subscriber*. Syllables with greater numbers of training examples are assigned more mixtures. It was empirically determined that one mixture would be assigned to the HMM for every 25 training examples for that syllable in the database. For the given training data, it was realistic to assign a maximum of 6 mixtures to each model. The original stepped mixture models included self-loops but no skips: skips were added to the models in a later experiment.

All experiments are repeated using syllable and word bigrams, derived from *Subscriber*, and insertion penalties at the recognition stage. In deducing the optimum insertion penalty, a series of recognition experiments are carried out on subsets of the database. The insertion penalty is varied until the number of insertions and deletions in the recogniser output are approximately equal.

In all cases, 16 cluster iterations and 20 estimation cycles are used for initialisation. Three cycles of embedded re-estimation are executed. A line noise model is also used, which has been previously trained on a range of non-speech sounds.

An example grammar for syllable recognition is given in Figure 2. Part of the syllable-delimited dictionary is shown in Figure 3, and the corresponding whole-word recognition grammar is given in Figure 4.

Further experimental details may be found in [7].

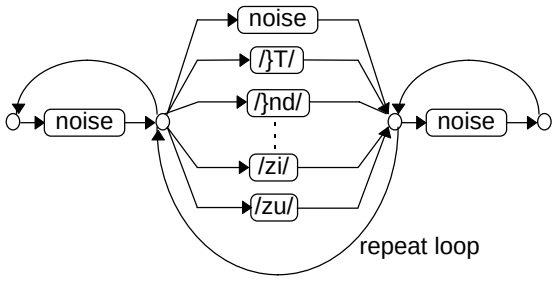

Figure 2: Grammar for unconstrained syllable recognition. This allows for any syllable in *Subscriber* to follow any other syllable. Provision is also made for line noise to precede or follow each syllable recognised.

```
A               eI
able            eI bl
zoo             zu
zuchini         zu ki ni
```

Figure 3: Part of syllable-delimited dictionary

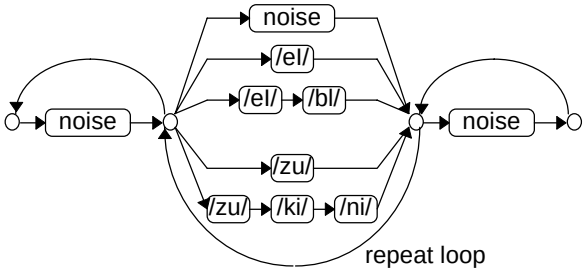

Figure 4: Corresponding grammar for word-constrained recognition using syllable models

3. RESULTS

The two tasks of unconstrained syllable recognition (Figure 2), and word recognition using syllable models (Figure 4), were tested for 1810 read sentences. Results are tabulated separately in Tables 1 and 2 respectively, the main results being shown in bold. The results in each table refer to the scores for the main unit to be recognised in each experiment - syllables in the case of unconstrained syllable recognition (Table 1), and words in the case of word constrained syllable recognition (Table 2).

There are two main recognition scores: the score correct and the score accurate. The score correct is simply the number of components (words, syllables or phoneme-like units) correctly transcribed. The score accurate is the score correct with 1 token subtracted for each incorrect insertion.

The corresponding results for monophone models are given in Table 3. These were accomplished using three mixture, no-skip models. It was unnecessary to include insertion penalties in this experiment, as there was already a greater percentage of deletions than of insertions.

Table 1: Unconstrained syllable recognition

Recog-nition unit	Bigram	Model topologies	Insertion penalty	% syll. acc.	% syll. corr	% syll. ins.
Syllable	**Y**	**One**	**Y**	**50.5**	**55.6**	**5.5**
Syllable	Y	One	N	36.5	51.3	14.8
Syllable	N	One	Y	12.4	17.9	5.5
Syllable	Y	Three	Y	39.5	43.4	3.9
Syllable	Y	Three	N	24.0	39.6	15.6
Syllable	N	Three	Y	15.2	20.8	5.6

Table 2: Word scores using syllable-based recognition models

Recog-nition unit	Bigram	Model topologies	Insertion penalty	% word acc.	% word corr	% word ins.
Word	**Y**	**Stepped**	**Y**	**61.1**	**71.2**	**10.1**
Word	**Y**	**Stepped**	**N**	**46.7**	**70.5**	**23.8**
Word	Y	One	N	46.7	69.9	22.2
Word	N	One	Y	16.7	26.3	9.6
Word	Y	Three	Y	43.1	55.8	12.7
Word	Y	Three	N	23.9	55.6	31.7
Word	N	Three	Y	14.9	25.2	10.3
Word	Y	Stepped, skip	Y	33.3	43.2	9.9

Table 3: Results for recognition using monophone models

Task	Bigrams?	Insertion penalty	% accurate	% correct
74-phoneme	Y	N	35.4	40.8
74-phoneme	N	N	27.0	34.5

4. DISCUSSION

As seen in Section 3, unconstrained syllable recognition, without using insertion penalties, gives syllable recognition accuracies of over 36%. Adding insertion penalties increases recognition accuracies, on the same task, to over 50%. Phoneme recognition using monophone grammars, with bigrams and without insertion penalties, yielded only 35% phoneme accuracy for the same test and training data - see Table 3.

It is appreciated that it may be slightly misleading to compare recognition using monophones, which contain no contextual information, with recognition using context-dependent units such as syllables. However, as explained in Section 5, it may be possible to build a recogniser that combines phoneme-like models, whole word models and syllable units. This could yield a further improvement in recognition performance, and will be the subject of further investigation.

The task of word constrained syllable recognition gives scores of over 60% word accuracy and 70% correct (as defined in Section 3). It is believed that this task better reflects practical applications of a syllable recogniser.

The one-mixture models generally show an improvement in recognition performance over three mixture models, suggesting that some three-mixture models are undertrained.

The most common syllable in the database occurs 3290 times in the training data. At the other extreme, 16 syllables had only 9 training examples. Figure 5 shows the proportion of syllables in *Subscriber* with given numbers of training examples.

An analysis of the recognition performance of individual syllables showed that 95% of the syllables with poor (less than 5%) recognition accuracy had less than 30 training examples. Figure 6 shows the variation of recognition performance with the number of training examples for each syllable.

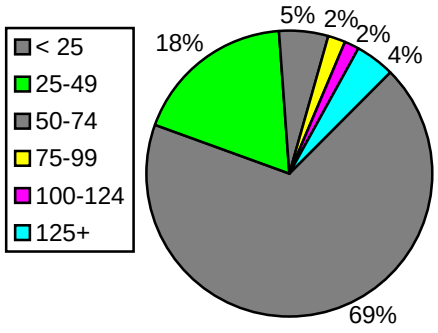

Figure 5: Frequency of training examples for syllables in training set

(Total number of different syllables: 1313)

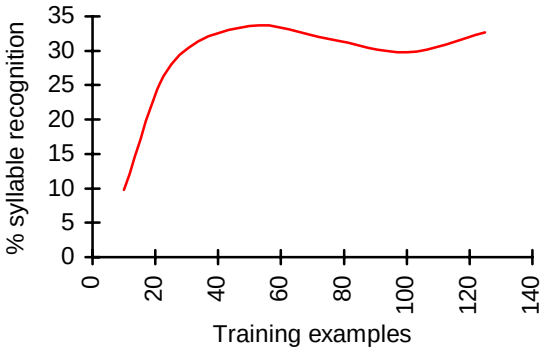

Figure 6: Recognition performance for syllables plotted against numbers of training examples

Predictably, then, the stepped mixture models, in which one mixture is assigned to the HMM topology for every 25 training examples for that syllable in the database, up to a maximum of 6 mixtures, give the best recognition performance of all.

A significant improvement is afforded by including word insertion penalties, due to the fact that there were a large number of insertions in the recogniser output. The insertions were mainly short-duration syllables, one or two phonemes in length. The most likely explanation of this is that the frame rate used in training and recognition was 16 ms, and the time-aligned transcriptions for *Subscriber* show that some syllables have durations as small as 4 ms.

Adding a skip to the model topologies gives a considerable decrease in recognition performance. This would seem to indicate that it is the process of forced recognition, used to generate the syllable-level time-aligned transcriptions, which gives rise to the inaccuracy in durational modelling.

The inclusion of bigram probabilities, derived from *Subscriber* rather than a more general language model, provides a significant increase in recognition performance. It should be noted, however, that 70% of the words in *Subscriber* and 29% of the syllables occur only in one context. Hence, including bigrams constrains the recogniser output significantly.

5. CONCLUSIONS AND FURTHER WORK

This work has shown that syllable modelling can substantially increase recognition performance on a medium-vocabulary database, when compared to monophone modelling.

Recognition performance is currently impaired by the large number of insertions in the recogniser output. Increasing the model insertion penalty alleviates this to some extent. Further experimentation is needed to investigate whether improved durational modelling would improve recogniser performance in this respect.

The question of out-of-vocabulary recognition has not yet been covered in the work presented here. This question becomes particularly crucial when considering that there are over 10,000 syllables in English [8], but that only just over 1,300 could be modelled using *Subscriber*. The out-of-vocabulary words could be modelled from another syllable-delimited database.

Currently, the experimental technique relies on dictionary transcriptions of all the sentences in *Subscriber*. This takes no account of the possible pronunciation variations within the database: some words in *Subscriber* were shown to have as many as 20 different pronunciation variants. To produce a transcription, by hand, of each utterance at the syllable level would be very time-consuming, but recent work at BT Labs has investigated the use of 'intelligent lexica' [9]. These use a set of alternative pronunciations and phoneme-based substitution rules, which have been shown to give a significant increase in recognition accuracy compared to an identical system relying solely on dictionary transcriptions. Experiments combining alternative pronunciations with the syllable models could be expected to yield a similar recognition improvement.

Syllable recognition implies increased computational costs and extra training data. Whereas phoneme-based modelling could be accomplished using approximately 70 HMMs, the number of syllable-level models required is loosely linked to the number of words in the training set. *Subscriber* contained 1243 different words, modelled using 1313 different syllable models. Smaller vocabulary databases, which have less instances of syllables occurring in more than one context or more than one word, would have a greater ratio of syllables to words.

These problems would be alleviated by a hybrid system where common words have corresponding whole-word HMMs, a full set of phoneme-based models is available, and the commonest syllables in the database are also modelled using HMMs. It is expected that this will be the subject of further experimental work.

6. REFERENCES

1. Hu, Zhihong; Schalkwyk, Johan; Barnard, Etienne and Cole, Ronald (1996): *Speech Recognition using Syllable-Like Units*, in *Proc. ICSLP 96*, Volume 2, pp. 1117-1120
2. Bourlard, Hervé and Dupont, Stéphane (1996): *A new ASR approach based on independent processing and recombination of frequency bands*, in *Proc. ICSLP 96*, Volume 1, pp. 426-429
3. Ladefoged, Peter (1975): *A Course in Phonetics*, Harcourt Brace Jovanovitch, New York, pp. 218-222
4. Laver, John (1994): *Principles of phonetics*, Cambridge University Press, pp. 517-518
5. Simons, Alison and Edwards, Keith (1992): *Subscriber - a phonetically annotated telephony database*, in *Proc. Institute of Acoustics*, Vol. 14 Part 6, Windermere 1992, pp. 3-15
6. Wells, John (1989): *Computer-coded phonetic notation of individual languages in the European Community*, in *Journal of the IPA*, Volume 19, pp. 32
7. Jones, Rhys J. (1996): *Syllable-based Word Recognition*, thesis for MSc Communication Systems, Department of Electrical and Electronic Engineering, University of Wales Swansea
8. Rabiner, Lawrence and Juang, Biing-Hwang (1993): *Fundamentals of Speech Recognition*, Prentice-Hall, p.436
9. Downey, Simon (1996): *Analysing alternative pronunciations to improve dictionary baseforms*, in *Proc. Institute of Acoustics*, Vol. 18 Part 9, Windermere 1996, pp. 331-338