MODELLING SYLLABLE CHARACTERISTICS TO IMPROVE A LARGE VOCABULARY CONTINUOUS SPEECH RECOGNISER

M. Jones & P.C. Woodland

Cambridge University Engineering Department,
Trumpington Street, Cambridge CB2 1PZ, UK.

ABSTRACT

The acoustic-phonetic modelling used in state-of-the-art large vocabulary continuous speech recognisers (LVCSR) cannot effectively exploit the prosody based distinctions known to exist at the syllable level. These distinctions are between the strength of the syllable (strong or weak) and the stress (stressed or unstressed) it is given. This paper shows how a small set of syllable-sized Hidden Markov Models (HMMs) can model syllable type effectively. These models have been applied to a large vocabulary continuous speech recogniser and a 23% reduction in word error rate was achieved.

1. INTRODUCTION

State-of-the-art large vocabulary continuous speech recognisers (LVCSR) achieve good performances using constraints including acoustic-phonetic models and language models. However, the search continues for additional constraints to reduce current levels of error and the likely greater error rates when speech complexity/variability intensifies, for example, in the case of spontaneous speech. Although, there is growing interest in higher level knowledge sources (KSs), such as sophisticated language and semantic models, it is important not to overlook lower level constraints.

This paper considers two syllable-based phenomena which have not been fully exploited in LVCSR systems. Syllables are speech units which extend over one or more phonetic segments: they have a nucleus (usually a vowel), preceded and succeeded by one or more consonants. Some work has been carried out on using syllable models as the basis of speech recognition but the concern, here, are general syllable characteristics: syllable strength (strong or weak) and stress (sententially prominent or not). Strong syllables are well formed over the entire syllable and in particular have full vowels. In contrast, weak syllables usually contain a schwa or a reduced vowel form. Strong syllables are also longer and louder than the weak ones. In general, syllables marked for stress in a lexicon are realised as strong syllables in speech; function words usually are realised with weak syllables. With every content word having at least one strong syllable, there are many strong syllables in any utterance. However, some of these will be distinguished further by pitch changes and greater loudness and duration: these are the stressed syllables.

The aim of the study was to see if the syllable characteristics could be modelled and then employed as additional constraints in state-of-the-art LVCSR systems. The KSs are attractive because of the low-level of the information being modelled: the syllable characteristics are carried directly in the speech, no sophisticated higher level processing is required and they can be modelled and employed in well understood ways.

A number of sets of syllable sized Hidden Markov Models (HMMs) were tested and the sets that gave the best classification used in an N-Best rescoring scheme. This was done in an attempt to improve the word error rate of a high performance LVCSR system for the TIMIT database.

The paper begins by describing the database used in all of the experiments. Then, the syllable modelling work is presented and the use of the characteristics in the LVCSR system is discussed. Experiments are reported that illustrate the problems of trying to model syllable characteristics directly in the front-end recognition system and which show the ways in which the N-Best approach can overcome these difficulties.

2. DATABASE

All of the experimental work was carried out using a LVCSR system for the TIMIT database. This database was chosen partly because of the availability pitch trackings for all of the utterances [6]. In order to have a task which was large but manageable, the speech material was was reorganised as in [3],[4]: only the *sx* utterances were used leading to a recogniser vocabulary of 1794 words; and, one third of all speakers were placed in the test set, two thirds in the training set (no speaker appeared in the test and training sets).

For all the data, time aligned syllable transcriptions were generated from the phone transcriptions. To do this, the TIMIT lexicon was used after it had been processed by a simple syllabification algorithm. The algorithm marked syllable boundaries and defined all syllables containing a lexically stressed vowel as strong and all other syllables as weak. All function words were marked as unstressed: in reality, some function words may be produced with stress for emphasis.

3. MODELLING SYLLABLE STRENGTH

The aim was to produce models that provided a good classification of syllable strength. To this end, the feature set, model inventory and the size of region modelled were considered.

3.1. Syllable HMMs

Strong/weak and stressed/unstressed syllables differ in respect to a number of features which interact in a complex way. Instead of attempting to build separate classifiers for the different features and combining these, the

HMM was chosen as the modelling tool. This work follows Freij and Fallside [1] who had success using HMMs to recognise lexical stress patterns in isolated words.

Five emitting states were used in each HMM with self and next-state transitions. Skip next-state transitions were also added to take account of variability such as the number of phones in pre-nucleic or post-nucleic clusters.

All the models were developed iteratively, beginning with a single Gaussian component per state and increasing the number up to 12 per state. The large number of components was required because of the broadness of the classes being modelled. As is described later, the models were applied given fixed boundaries: because of this a simple non-embedded training procedure was used.

3.2. Features

To capture the strong/weak distinction, the feature set should characterise articulation quality and loudness. To this end, spectral features (MFCCs) and normalised log energy were employed. First and second time derivatives of these features were also included.

To assess the usefulness of the different features in classification, models were trained and tested using various combinations of features (see Table 1).

Feature set	Features
Loudness	Energy + derivatives
Spectral	12 MFCCs + derivatives
All	Energy, 12 MFCCs + derivatives
Final	Energy, 6 MFCCs + derivatives

Table 1. Different feature sets used for the strong/weak syllable classifications.

It is well known that there are durational differences between strong/weak and stressed/unstressed regions. The standard HMM, however, models duration only in a very limited way. Relative duration models which accounted for different speech rates (following [4]) were built for the various syllable models but only insignificant improvements in classification were achieved due to large variances in duration. For this reason, no additional duration parameters were used in the final models.

3.3. Model inventory

In the first experiments just two HMMs were used, one for strong syllables and the other for weak ones. There was a wide range of syllabic structures, however, which were conflated by using the basic model set. For this reason, in a further experiment, four models were used for strong and weak cases, leading to a set of eight HMMs. The four models related to the syllable structures, CVC, CV, VC, and, V, where V was the nucleus and, C, a non-nucleic cluster of one or more phones.

3.4. Region modelled

An experiment was carried out to assess the effect of using just the vowel region for syllable classification. This involved building and training models on only the vowel sections of the strong and weak syllables.

In this experiment, the parameter set used was the *final* one (see Table 1) and the set of eight models were employed. Two types of model were tested: the first had five emitting states with the same transitions as the other syllable HMMs; and, the second used a phone-sized model structure of 3 emitting states with only self and next-state transitions. The latter set gave the best classification result of the two.

3.5. Assessing the models

As is explained in Section 7., the N-Best rescoring process did not require an unconstrained recognition of syllable types. Instead, syllable models were matched with acoustic vectors between given boundaries.

To assess the usefulness of the models for rescoring, the classification performance of the models given syllable start and end times was measured. For each test set syllable, the appropriate strong and weak models were used with the Viterbi algorithm to find the log-likelihood of the models accounting for the acoustic vectors. The best scoring model was chosen as the correct classification.

The different model sets tested and their performance on 500 of the test set utterances are shown in Table 2. Although all the different feature sets were tried for the last two model sets in the table, only the best results are shown.

Features	Num.	Reg.	%Err.S	%Err.W
Loudness	8	All	21	21
Spectral	8	All	24	25
All	8	All	15	17
Final	8	All	12	15
Final	2	All	22	29
Final	8	Vowel	24	29

Table 2. Different model sets tested and their performance on 500 test set utterances. Model sets differed with respect to the feature set, number of models (Num.) and region modelled (Reg.). Percentage strong (S) and weak (W) errors in classification are shown.

Although a good classification was possible with just 3 features (energy and its derivatives), the best performance was achieved with a combined feature set of energy, 6 MFCCs and first and second derivatives. The improved performance using 6 MFCCs over 12 was probably because the larger number of spectral features (useful for acoustic-phonetic decoding) have a high degree of variance through the broad nature of the data classes.

The best model set was the one with models for the different syllable structures modelling the entire syllable as opposed to just the vowel segment. Much better performance was seen than when just two models were employed due to the reduction of the variances of all parameters.

4. ADDING STRESS MODELS

The strong-weak model set was extended to try and exploit the additional distinctions between stressed-strong and unstressed syllables. The work involved adding a measure of pitch to the feature set and finding syllable candidates to train the stressed models.

Pitch was represented by normalised fundamental frequency ($f0$) values. These values were computed for every 10ms of each utterance. As a first step, the absolute fundamental frequency ($F0$) declination contour over the voiced portions of the utterance was approximated by carrying out a linear regression analysis. The mean voiced $F0$ value, $\overline{F0}$, was also found. Normalisation was then carried out by,

$$f0 = \frac{F0 - \hat{F0}}{\overline{F0}} \tag{1}$$

where $F0$ is the absolute $F0$ produced by a high quality pitch tracker [6], $\hat{F0}$ is the expected value from the declination line and $\overline{F0}$ the mean value in the utterance. In this way the normalisation took account of speaker differences and the $F0$ contour declination that is known to occur over the duration of statements.

Although it is a straightforward task to determine which syllables are likely to be strong or weak, the prediction of sentence level stress is not easy.

Ideally, for model training, hand labelled stress transcriptions would be available. However, for TIMIT this is not the case and is not realistic for any large task. Although, automatic prominence marking has been considered in [7] using parsing and additional prosodic information, in this work, a simple and robust approach to marking stress syllable candidates was devised.

The approach looked for strong syllables which had were likely to be stressed based on $f0$ and log energy values. In each training utterance, half of the strong syllables were marked as likely stress-strong. These syllables had the highest pitch and energy values in the utterance.

The extended model set consisted of 12 HMMs: four models for weak syllables and eight for strong syllables (4 strong and 4 trained on the stress candidates found by the above algorithm). For each syllable, the appropriate weak, strong and stressed-strong models were matched against the acoustic vectors and the syllable was classified by the model achieving the best log-likelihood score. In assessing classification performance, a strong syllable was scored as correctly classified if either the strong or stressed-strong models achieved the highest log-likelihood of the three models considered.

These new models improved strong-weak classification: 91% of strong syllables and 88% of weak ones were correctly classified. The average score difference between the correct and incorrect models was also increased.

5. BASELINE RECOGNISER

The work on syllable classification showed that there were quantitative differences between the syllable types. The next step was to try and use these differences to improve a TIMIT LVCSR recogniser.

The HTK toolkit [9] was used to produce an word-dependent N-Best recogniser [5]. This system was a state clustered word-internal triphone system with Gaussian mixture observation densities. The system was similar to the one described in [8]. There were 4065 triphones (1568 distinct states) with 3 Gaussian components per state (the optimal number). The parameters used in the system were: 12 Mel frequency coefficients, normalised log energy and the first and second derivatives of all parameters, giving a vector size of 39. The lexicon used was a slightly modified version of the one supplied with the database and the word-pair grammar was built from all the prompt material in TIMIT: the perplexity of the test set was found to be 41.

The same 500 randomly chosen test utterances used for the syllable classification experiments were employed to test the recogniser. The baseline performance on this set is shown in Table 3.

% Sentence error	16.2
% Word error	4.4

Table 3. Baseline recogniser performance for the 500 test set utterances.

For each utterance, the 15 best transcriptions were produced. Correct transcriptions were found for 93% of utterances. If all of the most accurate transcriptions were moved to the top of the N-Best lists, the word error rate of the recogniser would be 2%: this was the upper bound for improvements to be gained by using the new KSs.

6. PROBLEMS OF INTEGRATED MODELLING

It is difficult to integrate syllable characteristics with a front-end recogniser such as the one described above. Experiments were carried out to illustrate some of the problems.

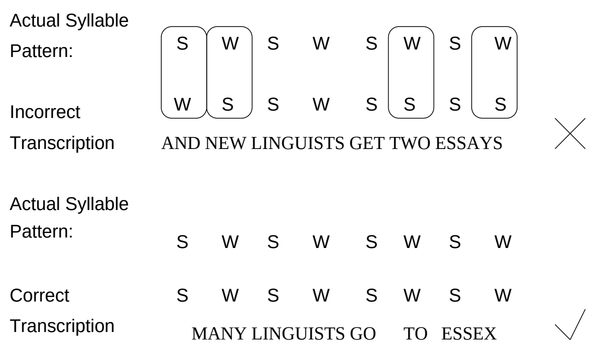

Figure 1. An example of scoring N-Best transcriptions with the strong (S) and weak (W) models. In the example, the incorrect transcription is scored with models which badly match the actual syllabic pattern (mismatches shown by boxes): this transcription scores poorly compared with the correct transcription.

One approach uses the syllable strength distinction as an additional context specifier when building context-dependent phone models: some researchers have achieved success with this technique. The triphone LVCSR, described above, was extended to take account of syllable strength. This involved first marking all lexically stressed content-word vowels as strong. A triphone inventory was then generated using this new lexicon leading to an additional 400 HMMs. The system was found to perform no better than the basic triphone one. Such context-dependent schemes, in fact, have several drawbacks: they model the strength differences over the vowel segment only, when, in fact, the distinction spans over a much greater region (see Section 3.4. for the effect of reducing modelling region); the amount of training data required increases dramatically (a problem in the TIMIT experiment); and, the parameters which can be used in the modelling are limited to those which are useful to acoustic-phonetic decoding.

An alternative scheme is to include additional parameters in the feature set used for recogniser. The problem with this approach is that the prosodic features cannot be easily or helpfully combined with acoustic-phonetic features. For example, a TIMIT phone recogniser using single Gaussian monophones was developed with the feature set consisting of 12 MFCCs, normalised log energy, $f0$ and first and second derivatives. Phone accuracy using this system was 45% this represented a 11% *increase* in errors relative to a recogniser which did not use $f0$ in the feature set.

7. N-BEST RESCORING APPROACH

In earlier work, the N-Best rescoring approach has been used to enable low-level KSs to be effectively integrated with a state-of-the-art recogniser ([3],[4]). Using the scheme enables different regions and parameters to be used than those that are effective for acoustic-phonetic decoding.

For every utterance, the N most likely hypotheses were found in order of decreasing log-likelihood and each had a log-likelihood score associated with it. The syllable models were then used to rescore N-Best transcriptions with the syllable boundaries being derived from phone transcriptions. The transcription syllable scores were also log-likelihoods and were combined with the front-end scores using weights optimised on the training set: the optimisation criterion was word accuracy. The N-Best list was then reordered on combined score.

For the strong-weak model set rescoring, all transcription syllables which belonged to content words and which

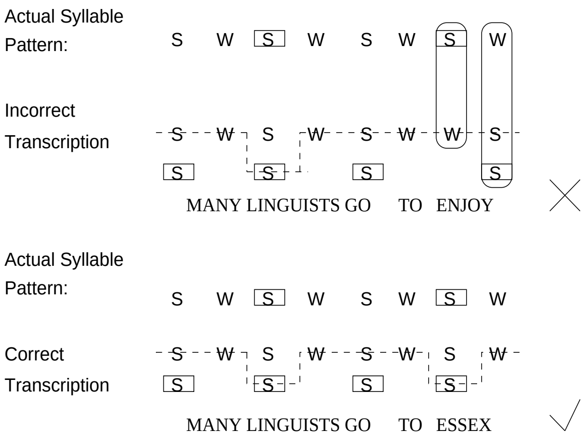

Figure 2. An example of scoring N-Best transcriptions with the extended model set. Stressed-strong models are boxed. Each proposed strong syllable is scored with strong and stressed-strong models: the best path through the set of models is shown as a dashed line.

were lexically stressed were scored using a strong syllable model and for all other syllables the weak model was used. In carrying out this scoring, the aim was to penalise transcriptions which proposed incorrect syllable types. An example of the scoring is shown in Figure 1.

The extended model set was applied in the same way as the strong-weak one except that strong transcription syllables were scored with two models, the strong but unstressed model and the strong-stressed one: the highest score was used in calculating the overall transcription score (see Figure 2). In this way, following [2], a loose stress constraint was enforced, penalising cases where a weak syllable was proposed and a stressed syllable occurred in the speech.

8. IMPACT ON BASELINE PERFORMANCE

The N-Best recogniser output was rescored with the strong-weak and extended model sets. The effects on the baseline performance are presented in Table 4.

Model set	% Word Err.	% Change
Strong-weak	3.7%	-16%
Stress	3.4%	-23%

Table 4. Effect of rescoring N-Best output with new model sets. The reduction in word error rates are shown relative to the baseline reported in Section 5.

Using the new model sets gave reductions in word error rate. This was possible because of the ways in which incorrect transcriptions were penalised relative to the correct or most accurate transcription in the N-Best list. Incorrect transcriptions scored less well than the correct ones for two reasons:

- **Mismatch on general syllable characteristics.** Transcriptions which proposed strong syllables where weak ones existed and vice-versa scored poorly. The extended model set gave a greater decrease in word error rate because of the additional discrimination between strong and weak syllables that it provided.
- **Mismatch on syllable structures.** Transcriptions which had syllabic structural patterns which were not the actual ones in speech were further penalised through the use of the syllable structure specific models. This type of constraint was seen to be useful in earlier work [3].

As an example, consider the top 2 transcriptions for the utterance, *"The big dog loved to chew on the old rag doll"*:

```
1 The|big|dog|loved   | to |an    |old rag doll
2 The big dog loved to chew on the old rag doll
```

The first, and incorrect transcription, scored worse than the second using the extended model set: the weak model used for "to" did not match well the actual strong syllable "chew"; and, at a number of points along the transcription the proposed syllabic structures conflicted with the actual ones, e.g., "an" with a structure of VC spanned a speech segment with the syllables "on the" (VC CV).

When considering the effect of the rescoring on the recogniser, it has to be remembered that the level of improvement possible is dependent on the quality of the N-Best output in the first instance.

9. CONCLUSION

This paper has shown how three classes of syllable can be modelled with a small set of HMMs. Ways of applying these models to improve the output from a recogniser that uses conventional knowledge sources have been presented. As with other studies, the work suggests that LVCSR systems should begin to include such prosodic KSs as additional constraints. The schemes presented, here, are applicable to state-of-the-art systems without being over complex or data-intensive.

ACKNOWLEDGMENTS

Matthew Jones holds an EPSRC studentship.

REFERENCES

[1] Freij, G. J. & Fallside, F. (1988). Lexical stress recognition using Hidden Markov Models. *Proc. ICASSP '88*, Vol. 3. 135-138, New York.

[2] Hieronymus, J. L., McKelvie, D. & McInnes, F. R. (1992). Use of acoustic sentence level and lexical stress in HSMM speech recognition. *Proc. ICASSP'92*, 225-227, San Francisco.

[3] Jones, M. & Woodland, P. C. (1993). Exploiting variable width features in large vocabulary speech recognition. *Proc. ICASSP '93*, Vol. 2. 323-326, Minnesota.

[4] Jones, M. & Woodland, P. C. (1993). Using relative duration in large vocabulary speech recognition. *Proc. EuroSpeech '93*, Vol. 1. 311-314, Berlin.

[5] Schwartz, S. & Austin, S. (1991). A comparison of several approximate algorithms for finding multiple (N-Best) sentence hypotheses. *Proc. ICASSP '91*, Vol. 1, 701-704.

[6] Tuerk, C. M. (1992). Automatic speech synthesis using auditory transforms and artificial neural networks. *PhD Thesis*, Cambridge University Engineering Department, U.K.

[7] Veilleux, N. M. & Ostendorf, M. (1993). Prosody/parse scoring and its application in ATIS. *Proc. ARPA Human Language Technology Workshop*, Feb. 1993.

[8] Woodland, P. C & Young, S. J. (1993). The HTK tied-state continuous speech recogniser. *Proc. EuroSpeech '93*, Vol. 3, 2203-2206, Berlin.

[9] Young, S. J. (1993). The HTK Hidden Markov Model toolkit: design and philosophy. Cambridge University Engineering Department Technical Report CUED/F-INFENG/TR.152.