EXPLOITING VARIABLE-WIDTH FEATURES IN LARGE VOCABULARY SPEECH RECOGNITION

M. Jones & P.C. Woodland

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

ABSTRACT

The use of variable-width features (prosodics, broad structural information etc.) in large vocabulary speech recognition systems is discussed. Although the value of this sort of information has been recognised in the past, previous approaches have not been widely used in speech systems because either they have not been robust enough for realistic, large vocabulary tasks or they have been limited to certain recogniser architectures. A framework for the use of variable-width features is presented which employs the N-Best algorithm with the features being applied in a post-processing phase. The framework is flexible and widely applicable, giving greater scope for exploitation of the features than previous approaches. Large vocabulary speech recognition experiments using TIMIT show that the application of variable-width features has potential benefits.

1. INTRODUCTION

For many years there has been an interest in the use of variable-width features (VWFs) in speech recognition systems. The aspects of interest include duration; the syllabic patterning in the utterance; the rhythm (the pattern of stressed/unstressed regions); and, the phrasing given by the prosodically marked boundaries. Despite this interest, the features have failed to be widely and fully exploited in large vocabulary continuous speech recognition (LVCSR) systems. In Section 2, the previous approaches proposed are reviewed with their drawbacks highlighted. The framework presented in this paper attempts to overcome these problems and to exploit the VWFs to a greater extent than in the past. The framework is outlined in Section 3. In developing the framework, a number of experiments were carried out. A baseline large vocabulary recogniser for the TIMIT database was built and is described in Section 4. Finally, the investigations illustrating the utility of the framework are presented in Section 5.

2. PREVIOUS WORK

A number of methods have been used to incorporate VWF information to aid the speech recognition task. For instance, VWFs have been used to drive the recognition process by identifying distinct/stable regions in the speech as anchor points for the decoding process [1]; to locate possible word boundaries [2]; and, to find sub-sets of the total recogniser vocabulary [3].

Another approach is to try to directly integrate the VWFs into the the recognition process. Attempts include the use of different subword unit models for different contexts (such as stressed/unstressed contexts [4]); the expansion of the feature set used in the recogniser to include such items as pitch (F0) and voicing [5]; and, complementing subword models with information from an independent VWF extraction process to constrain word-by-word lexical access [6].

A third way of using VWF knowledge is to post-process recognition hypotheses by adjusting the word recognition choices in some way. For example, [3] outlines an architecture using word intensities to rank hypotheses from a front-end recogniser and in [1] the recogniser grammar is modified to include phrase boundary information (based on F0) with word priorities being modified depending on whether they can come after a boundary.

A number of problems and limitations exist with all these approaches. Some methods that use the VWF knowledge to drive the recognition process are out of step with the state-of-the-art speech recognition methodologies. For example, using stressed syllables as starting points for recognition [1] belongs more to the earlier generation of knowledge based systems, where decoding proceeded explicitly step by step, than to the current stochastic based methods. With other methods there is the real danger of introducing errors (e.g. giving poor segmentation or possible word sub-sets) at an early stage before more useful and reliable constraints (such as phonotactic, lexical and grammatical ones) can be employed.

Approaches using the knowledge as recognition proceeds have several disadvantages. The effects we are interested in modelling often apply to larger contexts than single phone-like units; as many of today's recognisers use phone units as the basis of recognition, the integration of the knowledge and the models would seem to be a less than optimal. The domain over which the VWF knowledge can be applied may also be limited, (e.g., it is not feasible to make use of later parts of the utterance to modify the expectations about VWFs in the region currently being processed) and may be difficult to extend to new recognition schemes.

All previously proposed schemes for post-processing word choices are difficult to apply to the whole range of state-of-the-art recognition schemes and have features which make their suitability to LVCSR tasks questionable. For instance, in [3] word confusion lists have to be trained increasing the training data requirements.

3. THE FRAMEWORK

In developing a new framework the aim was to find a scheme which fits well with the state-of-the-art recognition methods and provides maximum flexibility for the effective use of VWFs.

3.1. N-Best Algorithm

In speech recognisers using Viterbi decoding the (single) most likely utterance transcription is produced. By contrast N-Best decoding algorithms generate the N most likely hypotheses. The sentence hypotheses are found in order of decreasing likelihood and each has a likelihood score associated with it.

An approximate, but computationally efficient N-Best algorithm, the word-dependent N-Best [7], has been implemented for a Hidden Markov Model (HMM) based recogniser. The HTK portable HMM toolkit was used as the basis of this recogniser [8]. For the framework to be useful, it is necessary to compute both the subword unit as well as the word segmentation. If the sub-word segmentation is not generated by the front-end recogniser, it is possible to find it by carrying out a second pass forced recognition using the word segmentation. Although we have used one particular form of HMM recogniser, the N-Best scheme can be applied to a range of state-of-the-art architectures.

3.2. Applying Variable-Width Features

The features are applied by looking for mismatches between the extracted VWFs and the features which would be expected for a given sentence hypothesis. Table 1 gives an example, with two possible sentence hypotheses for an utterance, along with their syllabic patterns. The utterance is *"I want to go to Birmingham"*; hypothesis 2 is the correct transcription, but the top hypothesis replaces "Birmingham" with "Brighton" (one less syllable).

For each transcription and each type of VWF (e.g., durational constraint, syllabic or stress patterns) a mismatch score is computed and combined with the initial front-end recogniser transcription likelihood score. The new overall score is used to reorder the N-Best list.

3.3. Framework Flexibility

There is a great deal of flexibility in terms of the way in which the knowledge can be applied using the framework. Mismatch scores can be calculated across an entire sentence hypothesis; alternatively, the match constraints can be relaxed: e.g., penalising a hypothesis if no variable-width feature is recognised at a point where, given the word segmentation, *at least one* would be expected. Features which have in previous work been used in higher level processes, can assist in penalising incorrect transcriptions: e.g., if a phrase boundary mark is found to coincide with the middle of a word in a sentence hypothesis, that hypothesis can be made less likely.

Using the knowledge in the proposed framework means that there is also flexibility in terms of how the variable-width features are extracted. It is not only possible to extract the features independently of the word recogniser output and then to match these against the each of the N-Best hypotheses but also to use the segmentation given by the recogniser to assist in the feature recognition process. For instance, instead of attempting to detect stressed regions in an unconstrained way across the entire utterance, the extraction process can examine syllable sized regions using the segmentation information from the recogniser. As much other work in the past has shown, extraction of features such as stressing is a difficult task: using the segmentation is one way in which our ongoing work will evolve, as we believe that it will improve the robustness and reliability of the extraction process.

	Transcription
Syllabic Pattern:	S - S - - - S - S - S - S — S — S - i—want—to–go–to–berm–ing–ahm
Hypothesis 1 Pattern:	S - S - - - S - S - S - - S - S i–want—-to–go–to—brai–ton
Hypothesis 2 Pattern:	S - S - - - S - S - S - S — S — S - i—want—to–go–to–berm–ing–ahm

Table 1. Syllabic patterns for utterance and 2-best hypotheses. "S" Syllable Nucleus; "-" non nucleus segment. Mismatch between hypothesis 1 and utterance pattern.

As entire sentence hypotheses are available before the knowledge is applied it is possible to make use of the features in ways which are not possible in other schemes. The expected feature patterns for a string can be determined using the sentence hypothesis (e.g., a method for predicting intonational boundaries from sentence texts is described in [9]) and the result of such an analysis scored against the actual boundaries found in the speech. Durational modelling is another area which can benefit from having the entire sentence hypothesis available. Although it is possible to model a variety of durational constraints within the front-end recogniser itself, it is not easy to take account of *relative* changes in duration caused for example by different speaking rates. Such sophistications are made possible in the post-processing scheme as all unit durations can be derived and the number of units making up the utterance is known. Using this information, a measure of the speaking rate could be found and the durational models used in rescoring adjusted appropriately.

Approaches which attempt to use the VWFs to drive the recognition process can place too much emphasis on the uncertain VWFs; in other approaches (e.g., [4]), the features seem to have too little impact. The N-Best scheme enables the emphasis to be adjusted in a formal way: the final order of the of the hypotheses can be a function of the different scores available, that is, of the front-end recogniser score for the hypothesis and the scores from applying various VWFs and the weights given to each score can be determined automatically [10].

Therefore, the proposed scheme is applicable to state-of-the-art recognisers and allows a great deal of flexibility in terms of the use of the features. Different features can be applied and extracted in different ways with their contribution to the overall ranking of hypotheses being made relative to other sources of knowledge.

4. EXPERIMENTAL SETUP

To test the application of the VWFs a number of large vocabulary speech recognition experiments have been performed. These experiments used the TIMIT database as it has a wide coverage of sentence patterns. In order to make the investigations more manageable, the total vocabulary was reduced. It was found that a good coverage of utterances can still be achieved even if the full vocabulary is substantially reduced from the total of 6229 words. For example, 2373 words account for all the test set utterances and just 600 words account for the 3150 *sx* type test utterances. From this analysis, a sub-set database was produced with a vocabulary of 1794 words and 450 different sentences (with multiple utterances from different speakers of each sentence). The new database was divided in a similar fashion to [11]: two thirds of all the speakers were placed in the train set and the remainder made up the test set (no speaker occurred in both the test and train sets).

A baseline speaker independent (continuous speech) HMM-based word recogniser was produced and used to process the test set sentences. A number of versions were built. Initially, 48 (10 mixture) HMM monophones, trained on the original TIMIT training set, were used. Word models were built using the pronunciations given in the TIMIT lexicon supplied with the database. The word error rate (w.err) of this version was 86.03% (no grammar). In the second version, the closure and release portions of stop phones were not modelled separately: compound stop models were produced. The pronunciation dictionary was also improved by using simple allophonic rules. These changes improved the performance to 83.01% w.err. The third version of the system (w.err of 75.75%) used the same model set/pronunciations as the second but the models were re-trained using phone-level transcriptions generated from the pronunciation lexicon. Finally a word-pair grammar (test set perplexity of 36) was added: this version had an error rate of 16.75% (word level) and 33% (sentence level).

The N-Best recogniser was used to generate 150 transcriptions for each of the utterances. Correct sentence transcriptions below the top of the N-Best list occurred for 15% of the test utterances. The majority of these transcriptions were found between the second and tenth position in the N-Best list.

5. USING THE FRAMEWORK

In developing the framework a number of investigations were carried out. The main work presented here involves the use of broad structural information in a post-processing phase; however, we also present some initial attempts at the use of relative duration information.

5.1. Broad Structural Information

Some preliminary work on a connected digit database had shown the benefits of using syllabic knowledge in post-processing the N-Best transcriptions. A small test set was used and 10-best transcriptions generated for each utterance. Knowledge about the location of syllable nuclei (identified by energy peaks in the signal) was applied and after reordering the list, the digit error rate was cut by 11% and the best transcription was found in the top 3 ranks in 25% more cases.

For the large vocabulary word recognition experiments using TIMIT, the syllabic pattern of the utterance was represented by sequences using just two symbols (NUCLEUS/BOUNDARY). Two types of investigation were carried out. Firstly, the actual broad representation of test set utterances were used: these were derived by taking the correct phone transcriptions for the utterances and replacing vowel phones with the NUCLEUS token and all other phones with the BOUNDARY symbol. These tests, described in (5.1.1) were designed to asses the framework's impact under a number of conditions. Then, attempts at directly extracting and applying the information were made (see 5.1.2).

5.1.1. Impact of Variable-Width Features

For each utterance in the test set, a broad representation of the 150 hypotheses was produced and the actual broad representation matched against each in turn. A new overall score was computed for each hypothesis which combined the original likelihood score and the broad representation mismatch score. When all hypotheses had been processed the N-Best list was sorted such that the top transcription had the highest overall score.

Initially, the match after a dynamic programming (DP) alignment between the expected and actual broad representation of a transcription was computed across the whole utterance. Table 2 shows the effect on system performance (relative to baseline) after post-processing for both the grammar and no grammar cases.

Condition	%change w.err	%change s.err
Grammar	-20%	-30%
No Grammar	-6%	-1%
Grammar + (a)	-15%	-18%
Grammar + (b)	-9%	-10%

Table 2. % decreases in word error rate $w.err$ and sentence error rate $s.err$ after application of VWFs.

To demonstrate the score-application flexibility of the framework, performance changes after two other methods were analysed: (a) using merged feature information and DP alignment in which adjacent broad class symbols of the same type were merged into one; and, (b) word match which checks to see if there is at least one vowel mark in the section of the broad transcription corresponding to each word in the N-Best hypothesis. Table 2 also contains the results of these tests. As well as improvements in terms word/sentence level accuracies, applying the features also caused correct transcriptions to move towards the top of the N-Best lists.

5.1.2. Extracting the Information

Attempts were made at directly extracting the features and then applying them as above. The "feature extractor" consisted of the 50 HMM phone models used in word recognition, mapped to just two output symbols (NUCLEUS/BOUNDARY). A phone bigram grammar was used as opposed to the lexically constraining word-pair grammar in the front-end word recogniser. This extractor had a 17.34% symbol error rate. This stochastic approach to feature recognition is one that our on-going research will focus on, using specific models of, for instance, stress and syllabic cues.

Various methods for compensating for the poor feature extraction were employed, including using score thresholds such that any score over the threshold was raised to the top score; restricting the variable-width feature application to the top i transcriptions; and, merging the features (as discussed above). The impact of applying the recognised features, with and without these compensating tactics are shown in Table 3.

5.1.3. Discussion

The results show that given a good recognition of broad structural information, significant improvements in recogniser accuracy and correct transcription position can be achieved. Top transcription accuracy is important whatever the application. The upward movement of transcriptions is of significance for spoken language systems: the nearer the top the transcription is, the less the likelihood of producing an incorrect system response.

The scoring-flexibility of the framework has been demonstrated, with improvements shown using different methods. The use of merged features, and the word-match scoring

Condition	%change w.err	%change s.err
No compensation	+168%	+158%
Threshold 60%, i=50	-4%	0%
Merged info. i=10	-6%	-7%

Table 3. %changes in error rates after applying extracted information. i is number of alternative transcriptions processed.

demonstrates that it is not necessary to extract every single feature (a difficult task to do, and a major problem with attempts to directly integrate VWFs into recognisers). The N-Best paradigm is amenable to the application of such a less-exacting representation of features whereas approaches which apply the features in a more limited context (such as word-by-word for lexical access) it is more critical to achieve fine, accurate recognition.

The poor performance of the feature extractor led to a significant reduction in effectiveness of the framework. Although simple tactics employed compensated to some degree there are additional methods we intend to use in future work: the different feature scores will be weighted (in the experiments, above, no such emphasis adjustments are made); and, instead of carrying out unconstrained recognition of the features, the segmentation given by the front-end recogniser will be exploited.

5.2. Relative Duration

Subword unit minimum and maximum duration constraints can be integrated easily into the front-end recognition process. We were interested in the use of such constraints adjusted for the speaking rate in the entire utterance (such adjustments are not easy to make as recognition proceeds).

Two classes of speaking rate, *faster* and *slower*, were defined where the classification was based on the average phone duration in the utterance. Phone duration minima and maxima statistics were derived from the training portion of the TIMIT database. As expected, the majority of phones in the slower utterances had higher minima and maxima than those in the faster ones. A single set of phone minima/maxima was also derived by analysing all utterances as one class: these are referred to as the standard statistics.

The correct transcription of an utterance should match the durational statistics better (or no worse) than the incorrect transcription. Transcriptions for two groups of utterance generated by the front-end recogniser were scored using the standard min/max information and also the speaking rate adjusted statistics. The percentage of utterances in each group where the correct transcription had a lower mismatch score than the incorrect one was found. The first group consisted of utterances that had a correct top transcription and an incorrect second place hypothesis. The second group had a correct second place hypothesis but an incorrect top transcription. For the standard duration statistics each transcription score was the percentage of phones in the transcription that had durations below the minimum or above the maximum observed in the training data. In the speaking rate adjusted case, the score was computed in the same way as for the standard case except that the set of durational statistics used was dependent on the speaking rate of the test utterance. The results are shown in Table 4. As can be seen, even when using a very simple adaptation to speaking rate, an improvement in the constraint modelling is achieved. In future work, more sophisticated durational constraints will be applied and adjusted using more flexible and complex measures of relative duration.

	Standard	Adjusted
Group 1	74.8%	77.4%
Group 2	44%	52%

Table 4. % of utterances in group where the correct transcription has a better match to durational constraints than the incorrect transcription.

6. CONCLUSIONS

The N-Best post-processing framework enables VWFs to be exploited in state-of-the-art large vocabulary speech recognition systems. Unlike previous approaches, the framework provides mechanisms to give the knowledge appropriate emphasis. The lack of robustness in some past schemes can also be overcome by virtue of the scoring flexibility inherent in the scheme and the use of front-end recogniser output to assist the feature extraction process. The framework also has the advantage of not being tied to a specific front-end recogniser architecture. The method presented allows the features to be used in new ways with, for instance, the availability of complete utterance transcriptions providing a useful additional source of information.

7. ACKNOWLEDGEMENTS

M. Jones is supported by a SERC Research Studentship. We would like to thank C. J. Leggetter for his work on the implementation of the N-Best algorithm.

REFERENCES

[1] Lea, W. (1980). Prosodic Aids to Speech Recognition. In Trends in Speech Recognition. Ed. W. Lea.

[2] Harrington, J., Watson, G. & Cooper, M. (1989). Word boundary detection in broad class and phoneme strings. *Computer Speech and Language*, **3**, 367-382.

[3] Waibel, A. (1988). Prosody and Speech Recognition. Pitman.

[4] Adda-Decker, M. & Adda, G. (1992). Experiments on stress dependent phone modelling for continuous speech recognition systems. *Proc. ICASSP'92*, 561-564, San Francisco.

[5] Robinson, T. (1991). Several improvements to a recurrent error propagation network phone recognition system. Cambridge University Engineering Dept., Technical Report CUED/F-INFENG/TR.82.

[6] 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.

[7] Schwartz, R. & Austin, S. (1991). A comparison of several approximate algorithms for finding multiple (N-BEST) sentence hypotheses. *Proc. ICASSP'91*, 701-704, Toronto.

[8] Woodland, P.C. & Young, S.J. (1992). Benchmark DARPA RM results with the HTK portable HMM toolkit. To appear *Proc. DARPA Continuous Speech Recognition Workshop*, Sept.'92, Stanford.

[9] Wang, M.Q. & Hirschberg, J. (1991). Predicting intonational boundaries automatically from text: the ATIS domain. *Proc. DARPA Speech & Natural Language Workshop*, 378-383, Feb.'91, Pacific Grove.

[10] Ostendorf, M., Kannan, A., Austin, S., Kimball, O. & Schwartz, R. (1991). Integration of diverse recognition methodologies through reevaluation of N-Best sentence hypotheses. *Proc. DARPA Speech & Natural Language Workshop*, 83-87, Feb.'91, Pacific Grove.

[11] Zhao Y., Waikita, H. & Xinhua, Z. (1991). An HMM based speaker independent continuous speech recognition system with experiments on the TIMIT database. *Proc. ICASSP'91*, 333-336, Toronto.