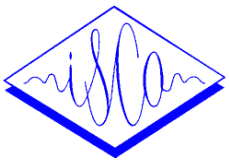

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

Modeling Sentential Stress in the Context of a Large Vocabulary Continuous Speech Recognizer

Kathleen Bishop

Dragon Systems, Inc.
320 Nevada St.
Newton, Massachusetts 02160

10.21437/ICSLP.1992-105

ABSTRACT

Recently, researchers have been studying the representation of "stress" and its relation to continuous phone recognition and continuous speech recognition. [4] has claimed an improvement in error rate due to an explicit marking of lexical stress when performing continuous speech recognition. On the other hand, [1] reported that using two levels of stress (as opposed to one level) did not reduce the error rate when performing phone recognition of continous speech. The English speaker-dependent continuous speech recognition system developed by Dragon Systems currently uses three levels of lexical stress for each of seventeen vowels. The English phoneme alphabet used by the system also includes twenty-six consonants (including three syllabic consonants, /M/, /N/, and /L/), totaling seventy-seven phonemes. The different stress levels lead to a large number of parameters to estimate when training models and performing recognition tasks. Motivated by the desire to downsize the parameter set, this paper is a preliminary study of how to maintain recognition performance when the number of stress levels is reduced. Performance results are reported in terms of the Wall Street Journal 5000-word vocabulary recognition task.

1. INTRODUCTION

This paper presents preliminary results of work done at Dragon Systems to improve recognition accuracy while reducing the number of parameters used in performing speech recognition tasks. The experiments have concentrated on different methods of modeling English sentential "stress". Currently the recognizer relies on three levels of stress, labeled on vowels: primary, secondary, and unstressed. The difference between primary and secondary stress can be heard when uttering the word "communicate". The second and the fourth syllable are both somewhat stressed when the word is uttered in isolation, but the second syllable seems more emphasized to the human ear, so it has primary stress. The fourth syllable in that word has secondary stress. Under the current system, the number of stress levels used in the lexicon influences the number of acoustic models that are built. This dependency will be explained further in Section 3. An explanation of the current modeling scheme for stress appears in Section 2, and a summary of the training data appears in Section 4. The experiments themselves are described in Section 5. The results are reported in terms of the Wall Street Journal 5000-word vocabulary recognition task.

2. MODELS OF STRESS

In this paper, lexical stress has been determined by statistical models that were seeded with hand-labeled data, and not by definitions of metrical phonology. For example, most phonologists would claim that the phoneme /æ/ does not have any reduced or unstressed forms. Nevertheless, the preliminary version of the system allows for an unstressed form of /æ/. This paper is a preliminary study of the usage of different numbers of stress levels for all vowels within a speech recognition system. Future work will include experiments that allow alternate stress levels for only those vowels that phoneticians claim to have alternate levels of stress.

In the current system, only vowels are modeled as having different stress levels, and three stress levels are used. Occurrences of stress are marked in the lexicon, reflecting the expected stress pattern of words spoken in isolation. These stress patterns are based on published general-use dictionaries' assignments of lexical stress. For a given word, alternate pronunciations that differ only in stress pattern may be included in the lexicon. Unfortunately for the recognizer, syllables that would be stressed when words are heard in isolation are not necessarily stressed when the same words occur in continuous speech. In other words, lexical stress (stress patterns of isolated words) and sentential stress (stress due to emphasis or phrasing) are not equivalent.

On the other hand, it may be reasonable to assume, as does [4], that any stressed syllable that occurs in continuous speech would be marked with either primary or secondary stress in the lexicon. For polysyllabic words this assumption sounds plausible, but if it is used the difficult choice arises of what stress level to assign to the vowels in monosyllabic words. One could mark the vowels in monosyllabic content words as being stressed, and the vowels in monosyllabic function words as having alternates in stressed and unstressed forms. This scheme would not lead to a large number of alternate pronunciations, but it would not allow for cases when a content word is reduced.

Letters provide a good example of the problems inherent in this approach. Lone instances of letters in continuous speech often have either primary or secondary stress. But when they occur together in abbreviations or other meaningful sequences, one or more of them might be considered reduced in some manner, like the "N" in the sequence "T.N.T.". Even though it is possible to treat letters as function words, and give them appropriate

alternate pronunciations, it seems to be an ad hoc way to model possible variation. A desirable scheme would be one that offers the recognizer some flexibility in its choices, but not by overloading it with many alternate pronunciations. Currently the system does give alternate pronunciations for monosyllabic function words, but all monosyllabic content words are marked to contain stressed vowels.

3. ACOUSTIC AND DURATION MODELING

The recognizer uses frame-based Hidden Markov Models (HMMs) for each phoneme, including explicit duration modeling probabilities for each state. Like other approaches that have appeared in the literature, the system uses triphone structures, called PICs (or phonemes-in-context), that can incorporate context information at the phoneme level.

Each PIC is represented by a linear HMM. The model for a particular hypothesis at recognition time is constructed by concatenating the necessary sequence of PICs, according to the specified pronunciation for each of the component words. This concatenation models both word-internal and cross-word co-articulation. The acoustic models may be shared across contexts for any given phoneme, but not across stress levels. At recognition time, if a PIC is required that has no existing model associated with it, a "backoff" strategy is called upon that uses a different but related PIC. Sometimes "generic" PICs are used in this case. These PICs are formed by calculating the modeling parameters for a given phoneme while allowing either the left side, the right side, or both the left and right side of the phoneme to vary across all possible contexts.

The duration distribution for each node of the HMM is assumed to have the double exponential form:

$$P(x) = \frac{1}{2\sigma} e^{-\frac{|x-\mu|}{\sigma}},$$

where μ is the mean and σ is the mean absolute deviation. Each node has an output distribution associated with it, based on thirty-two signal processing parameters: eight are spectral measurements, twelve are cepstral measurements, and twelve are cepstral estimated time derivatives. These output distributions are represented as thirty-two independent "streams" of data. Since it is well-known that the parameters within a frame are correlated, it may be preferable to group together the parameters into multi-parameter streams. Future work will include results that use different groupings of the parameters into streams.

The probability density of each stream is assumed to be a mixture distribution over a fixed set of basis distributions specific to the stream. This modeling paradigm, a variant of what is referred to as "Tied Mixtures" (TM) in the literature, is known to model well distributions that are multi-modal in character. For example, when all stress levels are merged into one level, (see Section 5) it is expected that this paradigm will be able to model the multi-modality of the resulting distributions of parameter values. [2] describes the implementation of TM in fuller detail.

Models for PICs are currently built starting with those for which there is the most data. Therefore, models for phonemes averaged over all left and right contexts are built first. Next, models for phonemes where the context is specified only on the right or on the left are built. Lastly the fully contextual PICs are built. When a relatively uncommon fully contextual PIC is built, models of related fully contextual PICs which share some of the context or have closely related contexts are used to smooth the model. At all stages, appropriate smoothing is done, using PICs that have been built at previous stages. [2] describes the training procedure in fuller detail.

4. DISTRIBUTION OF VOWELS

Table 1 gives the distribution according to stress of the five most frequent vowels, indicating the amount of training data available for each stress level. The table lists the vowels in IPA format. The number of different contextual models, or PICs, that were created for each of these vowels is listed in the third column, with the fourth column listing the number of training tokens available for that vowel. The second column gives an example of each vowel, with the letters in capitals indicating the syllable containing the vowel in question.

The table reveals that within a vowel the distribution in the training data according to stress was not uniform. In addition, across vowels, the relative frequency of each of the levels of stress varied. A major contributor to this non-uniformity was that half of the training data followed a verbalized punctuation format. About three hundred sentences for each speaker contained an utterance of "period", which accounts for the relatively high number of occurrences of unstressed [ʌ]. For the majority of the vowels, the amount of secondary stress training data available was much less than what was available for the other stress levels. Even though four of the five most frequent vowels have as much or more training data available for unstressed as for primary stress, for the other thirteen vowels that was not the case. Thus primary stress vowels considerably dominated the training data.

Vowel	Stress Level	Example	# PICs	# Tkns
[ɛ]	unstressed	as**SETS**	118	1004
[ɛ]	secondary	**REV**olution	140	1045
[ɛ]	primary	con**TEND**	339	12547
[i]	unstressed	per**I**od	691	15010
[i]	secondary	**RE**organize	94	53
[i]	primary	**PER**iod	569	11202
[ɪ]	unstressed	opera**TED**	436	23857
[ɪ]	secondary	**DIS**agree	163	1035
[ɪ]	primary	**IN**dian	355	10380
[ʌ]	unstressed	**A**sserted	731	26964
[ʌ]	secondary	some**BOD**y	117	534
[ʌ]	primary	**HUN**dred	233	8229
[ɝ]	unstressed	in**JURED**	791	12072
[ɝ]	secondary	net**WORK**	25	72
[ɝ]	primary	**DUR**ing	228	2933

Table 1: Distribution of training data for the five most frequent vowels over all speakers

5. EXPERIMENTAL RESULTS

This section compares different modeling schemes based on recognition performance on the 5000-word closed-vocabulary verbalized punctuation version of the Wall Street Journal (WSJ) task defined by NIST. The speakers spoke a sentence at a time, when generating the training and the test data. The training data can be split into two parts: about 300 sentences of verbalized punctuation data and about 300 sentences of non-verbalized punctuation data. The speakers pronounced all pronunciation marks as words, such as "period" for ".", when producing the verbalized punctuation data. The test data followed the verbalized punctuation format and varied from speaker to speaker. Each speaker recorded a different set of about 40 test sentences that were used for evaluating the different modeling schemes.

Word error rates for the 12 WSJ speakers are displayed in Table 2. The methods were seeded from models that were made in February 1992, and represent preliminary dry runs of the system. The baseline result, version 3S3D, was from February 1992. Whenever two stress levels were used, as in columns two and five of Table 2, the phonemes that were originally marked as being at a secondary level of stress were marked as unstressed. This partitioning was chosen over one that groups together secondary and primary stress because there were far fewer training tokens available for the unstressed vowels than for the primary-stressed vowels. (See Section 4). When two durations, but one stress level were used, as in column 4, the PICs that would have differed only according to stress in the two-level stress paradigm were given different duration distributions but they used the same output distributions. For the 2S1D paradigm, column 5, the PICs that differed only according to stress used the same duration distributions but different output distributions. The methods listed in Table 2, in order of appearance, are the three-level stress, two-level stress, one-level stress, one-level stress with alternate duration distributions, and two-level stress with no alternate duration distributions.

Spkr.	3S3D	2S2D	1S1D	1S2D	2S1D
001	5.8	5.9	6.9	7.6	6.6
002	6.5	5.4	6.3	7.0	6.1
00A	21.5	20.9	20.9	20.3	22.3
00B	19.0	19.5	20.7	19.1	17.0
00C	23.8	24.6	27.3	24.0	23.3
00D	23.1	21.9	22.7	21.9	20.9
00F	13.8	13.4	17.1	12.8	13.8
203	10.2	10.4	11.5	11.8	9.7
400	10.2	10.5	11.8	12.8	10.8
430	10.6	11.1	12.3	11.3	9.1
431	8.2	9.3	10.0	9.1	9.0
432	4.6	4.6	5.1	3.1	3.7
AVG	13.1	13.1	14.4	13.4	12.7

Table 2: Word Error Rate (%) for Different Stress Schemes

6. DISCUSSION

Overall, using only one stress level did degrade performance, but not by a considerable amount. The matched-pairs test from [3] was applied to appropriate pairings of the errors from the different methods, with the null hypothesis that there was no difference. Under this test, the 1S1D method was statistically significantly different from the 3S3D method (P value = 0.0002), the 2S2D method (P value = 0.0001), and the 2S1D method (P value = 0.00001). Therefore, the degradation in performance arising from collapsing to one level of stress is statistically significant.

Under the null hypothesis that there was no difference between the 2S1D method and the 2S2D method, the P value was 0.67. Under the test of the difference between the 2S1D method and the 3S3D method, the P value was 0.58. Thus, the slight gain in performance of the 2S1D models was not statistically significant. The important result here is that the collapsing of structure from 3S3D to only 2S1D certainly did not lead to any degradation in performance.

When the 1S2D method is compared with the 1S1D method, the P value is 0.01, so the performance improvement to the 1S1D method offered by adding alternate duration models is statistically significant. When the 1S2D method is compared with the 2S2D method, (P value = 0.11) and with the 3S3D method, (P value = 0.15), the performance degradation from collapsing to one level of stress but leaving alternate duration models is not as statistically significant. Lastly, the comparison of the 1S2D method with the 2S1D method (P value = 0.02), suggests that collapsing to one level of stress and leaving alternate duration models loses more information than removing all alternate duration models but allowing for two levels of stress. It is possible, however, that with better modeling of the duration parameters, the 1S2D method could improve its performance. The models for the speakers that had the worst word error rate, speakers 00A-00F, all benefited from the addition of alternate duration models to the one-stress scheme.

The most important results are firstly, that the error rates for the methods that used the most modeling information per PIC (3S3D and 2S2D) were not statistically significantly different from the method that used the least amount of duration information per PIC (2S1D). Secondly, even though the error rates for 3S3D and 2S2D were statistically significantly different from the methods that used the least amount of acoustic information per PIC, 1S2D and 1S1D, the degradations were relatively small, 2% and 10%, respectively.

7. CONCLUSIONS

Preliminary results show that the number of acoustic stress levels or the number of duration distributions available per PIC can be reduced, without significantly degrading the error rates, when performing recognition of read speech. These alternate schemes use fewer parameters on average and so result in considerable savings.

Future work will include the following experiments: (1) the addition of duration rules acting on consonants that are followed by stressed vowels, and not just on the vowels themselves, (2) improvements to the duration models, (3) the use of multi-parameter streams for the HMM output distributions, and (4) the use of alternate stress levels for only those vowels that phoneticians claim to have alternate levels of stress.

Finally, the original three-level stress method was a preliminary version that did not incorporate sharing of output distributions among PICs. Additional experiments will be done that compare schemes that are allowed to share output distributions across PICs.

ACKNOWLEDGMENTS

This work was sponsored by the Defense Advanced Research Projects Agency and was monitored by the Space and Naval Warfare Systems Command under contract N000-39-86-C-0307. The author would like to thank Francesco Scattone for help in setting up experiments, and John Bridle, Caroline Huang, and Barbara Peskin for their comments.

*

REFERENCES

[1] M. Adda-Decker, G. Adda, "Experiments on Stress-Dependent Phone Modeling for Continuous Speech", Recognition, *Proceedings of ICASSP*, San Fransisco, April 1992.

[2] J.K. Baker et al., "Large Vocabulary Recognition of Wall Street Journal Sentences at Dragon Systems", *Proceedings of the DARPA Speech and Natural Language Workshop*, Arden House, Harriman, NY, February 1992.

[3] L. Gillick and S.J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms", *Proceedings of ICASSP*, Glasgow, Scotland, May 1989.

[4] J.L. Hieronymus, D. McKelvie, F.R. McInnes, "Use of Acoustic Sentence Level and Lexical Stress in HSMM Speech Recognition", *Proceedings of ICASSP*, San Francisco, April 1992.

[5] D.B. Paul, "New Results with the Lincoln Tied-Mixture HMM CSR System", *Proceedings of the DARPA Speech and Natural Language Workshop*, Pacific Grove, California, February 1991.

[6] A. Waibel, "Suprasegmentals in Very Large Vocabulary Isolated Word Recognition", *Proceedings of ICASSP*, San Diego, California, March 1984.