USING STRESS INFORMATION IN LARGE VOCABULARY SPEECH RECOGNITION

Pierre Dumouchel, Matthew Lennig*
INRS-Télécommunications (Univ. du Québec)
3, Place of Commerce
Verdun, Québec CANADA H3E 1H6

Using stress information in a Markov source-based large vocabulary speech recognition system provides a way to examine a nonlocal cue which is generally poorly represented by the Markov source model. In this paper, we present an algorithm for estimating the stress pattern based on syllable durations and short-time energies. The output also gives the probability of the correctness of the estimated stress pattern. The parameters are first normalized in an attempt to reduce variability due to different linguistic contexts. The stress pattern is then estimated based on a statistical approach. After initial training, tests on a new word list yielded 95% correct detection of the syllable carrying the primary stress. Finally, inclusion of this algorithm in a large vocabulary isolated word recognition system contributes to its accuracy.

INTRODUCTION

The goal of this research is to devise an algorithm for the estimation of the stress pattern of a spoken word from its acoustic signal. Such an algorithm would serve as a component of a speech recognition system. Input to the stress pattern estimation algorithm consists of a word's hypothesized phonemic transcription with stress markers and the corresponding acoustic signal. The output is a probability estimate of the correctness of the hypothesized stress pattern assuming the segmental transcription is correct. Only duration and short-time energy are used as parameters.

The definition of stress differs depending on whether we regard it from the point of view of the speaker or from the point of view of the hearer. From the speaker's standpoint, stress may be defined in term of greater effort to produce a syllable. From the listener's standpoint, stress is manifested by duration, energy level and increased (or decreased) pitch. Moreover, stress information is not strictly localized in time but requires information from the surrounding syllables of the word. In other words, stress is a contrastive nonlocal cue which overlaps adjacent segments because it is expressed *relative to other segments*. In this work, since we are interested in speech recognition applications, we will adopt the listener's point of view.

The purpose of the stress pattern estimation algorithm is to sharpen the overall accuracy of a Markov source-based speech recognition system. The incorporation of this algorithm as a module in a recognition system will also provide a way to examine a contrastive nonlocal cue. Nonlocal cues are poorly represented in the framework of Markov models.

A published lexical stress detection technique due to Aull (1984) may be described as categorical since no confidence estimate of the decision correctness is made. Aull tries to find the primary stress syllable of the word and the remaining syllables are labelled by rules as either unstressed or reduced. The present paper explores a probabilistic lexical stress detection technique. First, a normalization is applied on the energy and duration parameters in an attempt to reduce the variability due to different linguistic contexts. Second, the algorithm finds the hypothesized primary stress syllable based on a statistical approach. Finally, the probabilities of the estimated stress pattern and the lexical stress pattern (as given by the Webster's Seventh New Collegiate Dictionary) are evaluated.

* also with Bell-Northern Research, Montréal, Canada

DESCRIPTION OF THE ALGORITHM

Duration, energy level and pitch of the syllable are phonetic correlates of stress. But stress is not the only phenomenon which exerts an influence on these parameters. Intrinsic phonetic characteristics, phonological context, prepausal lengthening and speaking rate may also affect them. Hence the lexical stress algorithm uses a series of fixed correction factors to compensate for each of these effects except stress. In this study, only the duration and energy level cues are used. Pitch is not employed due to the difficulty of extracting reliable fundamental frequency information. Since stress principally affects the vowel part of the syllable, we judge it to be sufficient to examine only this class of phonemes. By doing so, we avoid having to segment difficult classes of phonemes such as initial and final voiceless stops. Hence, the duration cue used by the stress algorithm is the duration of the vowel part. Similarly, the energy level cue is the average of energy level over the vowel. The phonemic segmentation is based on a Viterbi alignment technique.

Normalization of intrinsic phonetic characteristics is used to compensate for the intrinsic duration and intensity of the vowels. For example, for the same source power the high-front vowel i will generally be less intense than a low-back ɑ . Hence, compensation factors are proposed for the intrinsic phonetic characteristics to counter this variability. The compensation factors that we use come from two studies of Lehiste (1960, 1970). Similarly, phonological normalization is used to compensate for the influence of the adjacent phonemes on the duration of the vowel. For example, a vowel is longer if the syllable ends with a voiced stop rather than a voiceless stop. The phonological duration compensation factors come from the previously cited Lehiste study (1960). The phonetic description of the syllable is given by the dictionary. No phonological energy compensation factors are proposed. A fixed factor is proposed to compensate for prepausal lengthening. Finally, linear normalization of parameters, such that within a word the normalized durations sum to unity and the normalized energies sum to unity, acts as a compensation for speaking rate and overall speaking level effects.

Figure 1 shows the distribution of vowels based on stress type for a corpus of 135 two-syllable words read in isolation from a text by a male speaker. The symbol P stands for *primary stress*, S for *secondary stress*, U for *unspecified* stress as given by the dictionary. The unspecified stress syllable is one with no lexical stress marker and it corresponds either to a ternary stress syllable or an unstressed syllable. The vowels are represented by their normalized, compensated parameters. No evident demarcation between the unspecified and secondary stress syllables is seen. It appears from this figure that three regions can be identified: a region where the primary stress syllables predominate, another one where the unspecified stress and secondary stress syllables predominate, and finally an overlapping region where all the

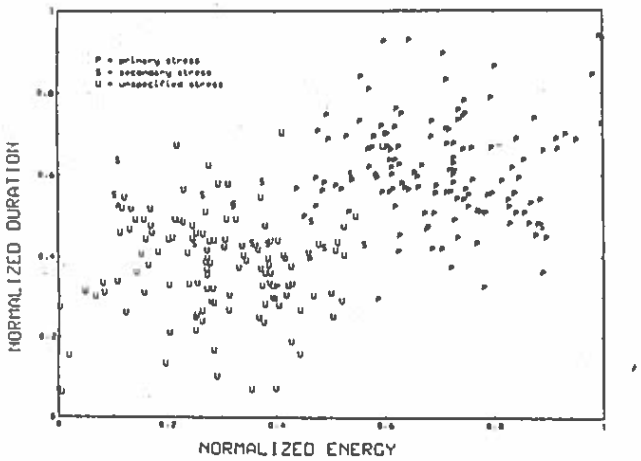

Fig. 1 Distribution of vowels.

types of stress are present. Similar figures are obtained for three and four syllable words but with different centers of gravity for each region. The difference between centers of gravity is due to the fact that we normalize the sum of each parameter to unity regardless of the number of syllables in the word. We conclude that a statistical approach is viable only to differentiate between the primary stress syllable and the other types of syllables (including the secondary and unspecified stress syllables). Furthermore, it appears that an additional normalization factor applied to each parameter for words containing more than two syllables can produce plots with centers of gravity similar to two-syllable ones. Based on these facts, the energy-duration space has been partitioned into 41 regions. The regions are enclosed by straight lines with slopes of minus one. Regions corresponding to the overlapping region are of smaller dimensions to achieve finer discrimination at the category boundary. We allocate to each region a probability denoting a specific type of stress. The probability is based on the frequency of appearance of a specific type of stress within a region compiled from a list of 220 polysyllabic words:

$$\Pr[stress = X \mid region = Y] = \frac{number\ of\ X\ in\ Y}{total\ number\ in\ Y}$$

A hand-smoothed version of results obtained with the above equation has been used. This is necessary to avoid unwanted effects of the relatively small size of the corpus such as a region not containing any data points. Finally, the estimated primary stress for the word is assigned to the syllable which has the highest probability of being primary. The probability of the maximum likelihood stress pattern is estimated as the average of syllable probabilities with respect to its estimated type of stress. We use the average of syllable probabilities instead of the multiplication of syllable probabilities since the latter incorrectly favors words with the smallest number of syllables. The lexical stress probability is determined in a similar way except the stress pattern is now the one proposed by the dictionary.

RESULTS

After initial training, tests on the same speaker reading a new word list yielded 95% correct detection of the primary stress syllable when compared to the lexical stress pattern. A list of 50 words pairs such as *PERfect—perFECT* (noun/verb) and a 220 polysyllabic words constitute the training word set. The test set contains 112 new polysyllabic words. The syllable distribution within the test corpus is the following: 66% two-, 23% three-, 10% four- and 1% five-syllable words. An examination of the errors reveals that of the 5% errors, three-fifths are due to incorrect phonemic segmentation produced by the Viterbi algorithm and one-fifth are due to a stress pronunciation of the word which differs from that of the dictionary. A final test which consists of examining the contribution of this algorithm in a large vocabulary speech recognition system has been performed. The recognizer uses hidden Markov models to hypothesize a list of words with their associated probabilities. During this test we modify the likelihood of each word derived from the acoustic data by the probability that the required lexical stress pattern is supported by the observed stress data. Results show that the rank of the correct word in the word hypothesis list improves by an average of 0.3 word positions when using stress information. This test is performed using 60 test words. However, for two-thirds of the list the correct word is already ranked first, so no improvement is possible. Excluding these top rank cases, the improvement amounts to an average of 0.9 word positions.

DISCUSSION

Lexical stress can be useful in recognition but its estimation is difficult because

- even in isolated word speech, word stress differs from the lexical stress pattern (1% of cases),
- the lexical secondary stress syllable is considered less stressed than the unspecified stress syllable of the same word in 30% of cases, based on a perceptual experiment with one subject on a list of 25 words.
- normalized duration and short-time energy parameters for secondary and unspecified stress form overlapping classes.

Hence an approach which attempts to find the primary, secondary and unspecified stress syllables of the word is excluded. However, an approach which consists of finding only the primary stress syllable is possible and can also offer a good constraint. By expressing the confidence of the detection probabilistically, the performance of the algorithm can be integrated with the results of the other recognition system modules. The technique described in this paper respects these constraints and the performance of the algorithm is extremely satisfying. However, the contribution of the algorithm to a large vocabulary speech recognition system is only a minor improvement in the rank of hypothesis. Further improvements are anticipated from a better match between relative likelihoods based on acoustic-model estimation and stress estimation.

REFERENCES

Aull, A.M., *Lexical stress and its application in large vocabulary speech recognition*, Master's thesis, Massachusetts Institute of Technology, 1984.

Peterson, G.E. and Lehiste, I., "Duration of syllable nuclei in English", *Journal of the Acoustical Society of America*, vol. 32 no. 6, June 1960.

Lehiste I., *Suprasegmentals*, The MIT Press, Cambridge, Massachusetts, Chapter 4, 1970.