MODELLING THE PROSODY OF SIMPLE ENGLISH SENTENCES USING HIDDEN MARKOV MODELS

Michael Wagner, Bob McKay, Santha Sampath, David Slater

Department of Computer Science
University College/ADFA, University of NSW

A set of 144 declarative sentences with a subject-verb-object structure is drawn from a vocabulary of monosyllabic and disyllabic English words. Fundamental frequency contours and energy contours of the sentence set are analysed with respect to the syllabic structure of the sentences. Multivariate correlation analysis provides predictions for the average energy and fundamental frequency of syllables. Based on the distributions of the energy, voicing and fundamental frequency parameters, 2 different continuously variable Hidden Markov Models are trained to distinguish between intersyllable intervals, stressed and unstressed syllables. One HMM uses single-mixture Gaussian parameter distributions while the other uses double mixtures. The Viterbi algorithm is used for automatic segmentation. It is noted that the convergence of the training procedure is sensitive to the initial distributions. It is also argued that the inclusion of duration modelling is essential to distinguish between stressed and unstressed syllables.

1. INTRODUCTION

The potential contribution of prosody-related speech parameters to automatic speech recognition has long been recognised. Yet, there is still little hard evidence as to how energy and fundamental frequency contours including their timing characteristics can be utilised in automatic speech recognition systems.

In this context, the question arises whether it is possible to derive the stress pattern of a speech utterance from the prosodic parameters that can be measured from the speech waveform. Such information, related to either the word stress or the sentence stress patterns could then conceivably be integrated into an automatic speech recognition system in order to achieve higher recognition rates for isolated-utterance or continuous-speech recognition.

Two experiments were conducted to determine the extent to which stress patterns can be modelled using energy, fundamental frequency and voicing measurements. In the first experiment, a small corpus of speech data, containing simple subject-verb-object sentences was recorded and analysed with respect to the average energy and fundamental frequency for each syllable. The speech data is described in detail in Section 2 and this experiment itself is reported in Section 3.

The second experiment uses the same corpus of speech data to derive the statistical distributions of the three acoustic parameters energy, voicing and fundamental frequency. Two different Hidden Markov Models were then trained to recognise the three states "stressed syllable", "unstressed syllable" and "intersyllable interval". The results of this experiment are reported in Section 4.

2. SPEECH DATA

The speech material is restricted to simple transitive sentences, with one-word subject noun phrases followed by one-word verbs and one-word object noun phrases. The words chosen are either monosyllabic or disyllabic, the latter class having a stress on either the first or second syllable. Some care was taken to ensure that both relatively high and relatively low vowels were represented in the words chosen in order to balance the intrinsic effects of vowel quality on F0.

Four words were chosen for each of the monosyllabic (11-words), disyllabic with stressed first syllable (12-words) and disyllabic with stressed second syllable (22-words) classes, for both nouns and verbs. This gives a total of twelve nouns and twelve verbs. The words were chosen so that half of the words in each class had relatively high vowels and the other half had relatively low vowels. The words chosen

are listed in Table 1.

Each of the twelve nouns was used as a subject noun phrase with each of the twelve verbs. This produced a set of 144 sentences. Each of the twelve nouns was also used twelve times as an object noun phrase. The order of the sentences, and the object noun phrase for each sentence were determined pseudo-randomly. A selection-with-replacement algorithm was used which allowed the same noun to be both the subject and object in a single sentence, and this in fact occurred in five of the sentences.

	Nouns		Verbs	
	High vowel	Low vowel	High vowel	Low vowel
11-words	Greeks kings	mobs tarts	beat leave	mark rob
12-words	teachers students	farmers robbers	nibble tutor	charter honour
22-words	cadets patrols	Pathans Malays	admit respect	garotte retard

Table 1. Word selection for the sentence generator.

The sentences were recorded by a male speaker of Australian English and digitised at 8000 samples/s and 12 bits/sample. The signal energy was determined for each 16ms frame and the voicing and fundamental frequency parameters were determined every 16ms with a centre-clipping autocorrelation algorithm based on a frame size of 48ms.

Syllables were marked automatically with some manual corrections. For each syllable, maximum, average and total energies, average and central fundamental frequencies, zero-crossing rate and first linear-prediction coefficient were determined. In addition, the maximum-energy frame for each syllable was recorded for the determination of syllable and stress timing.

These acoustic parameters were then correlated with the different structures of the recorded sentences.

3. ENERGY AND FUNDAMENTAL FREQUENCY AND VOICING STATISTICS

In the first experiment, the dependencies of the average energy and fundamental frequency of a syllable on the syllabic structure of the sentence were investigated. As each sentence of the recorded material consists of subject, verb and object, and each of the three words can be monosyllabic or disyllabic with word stress on either the first or the second syllable, the recorded material was divided in 27 groups of sentences where each group has a characteristic syllabic structure.

The three different word structures are denoted as "11" for a monosyllable", "12" for a disyllable with word stress on the first syllable, and "22" for a disyllable with word stress on the second syllable. This notation is extended to sentence syllabic structures in the obvious way. For example, "11-12-22" denotes the syllabic structure of the sentence "Kings honour patrols", etc.

The results of the analysis for the average fundamental frequency of the syllable are shown in Figure 1. The horizontal axis represents the sequence of stressed (fat dots) and unstressed (hollow dots) syllables for the 3 words. The vertical axis represents the average fundamental frequency for the given syllable over the group of sentences with the given structure.

For most of the 27 different sentence structures, the falling tendency of a fundamental frequency contour from the first to the second word is easily observed. From the second to the third word, the speaker tends to raise the fundamental frequency very slightly. Within disyllabic words, most transitions between the two syllables are marked by a distinct fall in fundamental frequency.

Multivariate correlation analysis shows the following correlations between fundamental frequency, position of word in the sentence and syllable structure of the word:

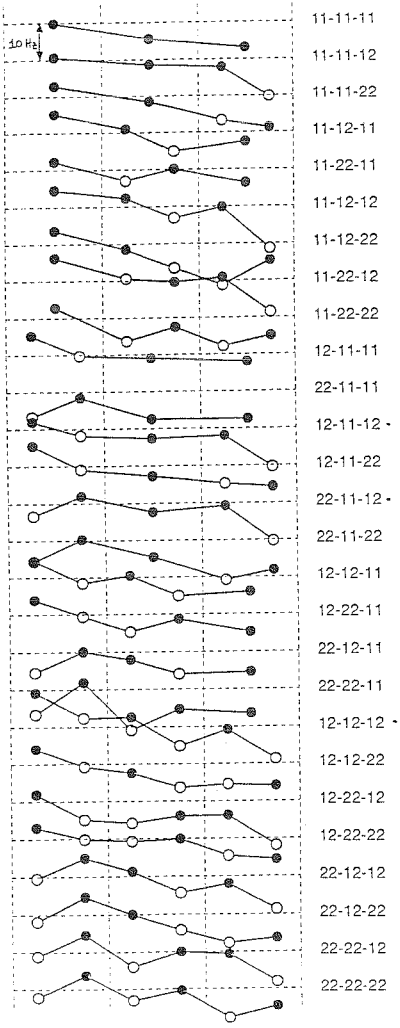

Figure 1. Average syllable F0 vs syllabic structure.

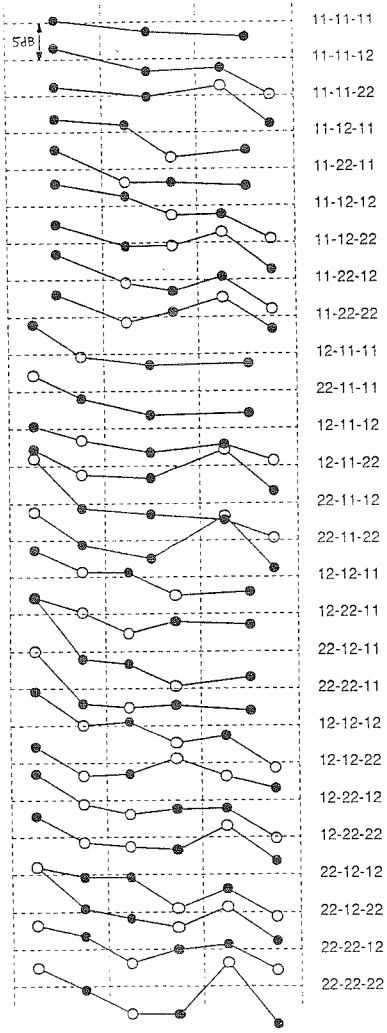

Figure 2. Average syllable energy vs syllabic structure.

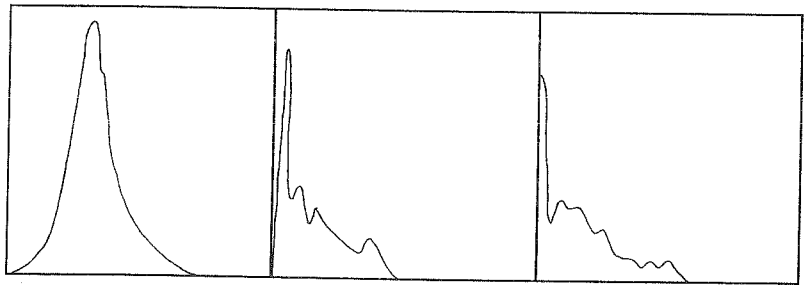

Figure 3. Energy, voicing and fundamental frequency distributions for State1.

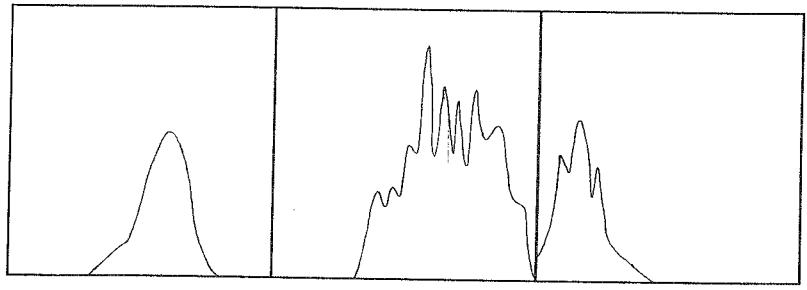

Figure 4. Energy, voicing and fundamental frequency distributions for State2.

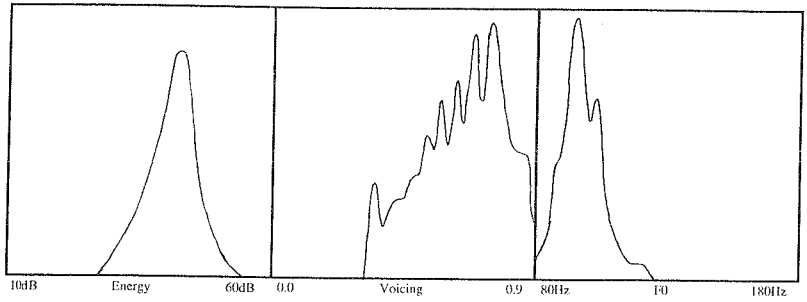

Figure 5. Energy, voicing and fundamental frequency distributions for State3.

Introducing the variables x_1 for the position of the word in the sentence (x_1=1,2 or 3), and x_2 for the word structure (x_2=0 if structure=11, x_2=-1 if structure=12, x_2=1 if structure=22) the average fundamental frequency for the syllable may be expressed as

$$F0_{ave} = 106.4Hz - x_1 * 3.7Hz + x_2 * 0.7Hz$$

with x_1 accounting for 18% and x_2 for 1% of the variance of $F0_{ave}$ and a total correlation of R=0.44.

The corresponding average energy contours for the 27 sentence categories are shown in Figure 2. In this figure, the vertical axis represents the average energy over the sonorant part of the syllable measured in dB.

The figure shows a decline of syllable energy over the sequence of words which corresponds to the falling tendency of the fundamental frequency shown in Figure 1. The other dominating influence on the syllable energy which is clearly shown is of course the higher energy for stressed syllables as compared with the unstressed syllables.

Modelling the energy contour on the variables x_1 as defined before and x_3 for the word stress (x_3=1 for a stressed syllable, x_3=0 for an unstressed syllable), results in the model for the average syllable energy of

$$E_{ave} = 43.4dB - x_1 * 1.7dB + x_2 * 2.3dB$$

with x_1 accounting for 25% and x_2 for 15% of the variance of E_{ave} and a total correlation of R=0.63.

The existence of this correlation between stress patterns and the energy and fundamental frequency parameters provided the motivation for a second experiment.

4. AUTOMATIC SYLLABLE SEGMENTATION USING CONTINUOUS HMM

In the second experiment the distributions of the three prosodic parameters energy, voicing and fundamental frequency were determined on a frame-by-frame basis for stressed syllables, unstressed syllables and intersyllabic intervals. On the basis of these distributions two different Hidden Markov Models were trained, one using single Gaussian parameter distributions, the other using mixture-Gaussian parameter distributions.

Figure 3 shows the distribution of frame energy, voicing and fundamental frequency for stressed syllables (State 1), Figure 4 shows the corresponding distributions for unstressed syllables (State 2) and Figure 5 shows the corresponding distributions for intersyllable intervals (State 3).

The energy distributions show a clear trend towards higher energies as one progresses from State1 to State3 with the mean values of 27dB for State1, 39dB for State2 and 41dB for State3. However, it is also clear that the distributions overlap considerably, especially the unstressed vs stressed syllable distributions.

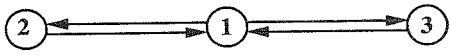

Figure 6. 3-State Markov Model for Intersyllables (1), Unstressed (2) and Stressed (3) Syllables.

The distributions of the voicing parameter, shown in Figure 4, are significantly wider than the energy curves while also showing an increasing tendency from State1 to State3 with mean values of 0.11 for State1, 0.54 for State2 and 0.57 for State3.

The fundamental frequency distributions, shown in Figure 5, are virtually identical between State2 and State3 with mean values of 96Hz and 98Hz while the distribution for State1 is entirely due to the behaviour of the autocorrelation algorithm for unvoiced speech segments.

Figure 6 shows the Markov Model which represents the state transitions over the course of the sentences. All transitions with the exception of transitions between states 2 and 3 are allowed as the system must enter an intersyllabic interval in between each two syllables. Initially, the system is forced to be in State1.

The means and variances of these distributions formed the basis of the training of the two Markov models. For the single-mixture model, the means and covariances as determined previously were used as the initial model. For the double-mixture model, reasonable means and covariances for the initial model were estimated graphically. In the double-mixture model, a diagonal covariance matrix was used for the initial model.

The forward-backward algorithm and the Baum-Welch reestimation procedure were used with the first 72 sentences of the recorded material for training until convergence of the model parameters could be observed after 20 iterations. The Viterbi recognition algorithm was then used to recognise the sentence structures of the remaining 72 sentences and to segment them into sequences of States 1, 2 and 3.

Table 2 shows the confusion matrices for the frame labelling using the Viterbi algorithm. Table 2a shows that the single Gaussian pdf results in an overall correct frame recognition rate of 75.7% with most of the intersyllables recognised correctly but with significant confusion between stressed and unstressed syllables. The model in fact labelled the majority of the unstressed frames as stressed.

Table 2b shows that the assumption of a double Gaussian mixture pdf in conjunction with diagonalised covariance matrices worsens the automatic segmentation results to 61.4%. A majority of intersyllables are now labelled as unstressed syllables. Careful analysis of the convergence of the model during the training phase leads to the conclusion that the model converges away from the means initially estimated from Figures 3, 4 and 5 towards a configuration in which State2 attracts more and more of the observations from both State1 and State3.

We argue that the training of this Hidden Markov Model is very sensitive to the initial statistical distribution of parameters, and that the diagonalisation of the covariance matrix leads to the convergence of the Baum-Welch algorithm towards a suboptimal maximum.

Furthermore it is evident from the two experiments reported here in conjunction with a recently published analysis of syllable timing that the inclusion of duration timing in the Markov Model is essential for the reliable modelling of stress patterns in speech.

		State Determined by Viterbi		
		State1	State2	State3
Actual State	State1	2886	41	139
^	State2	103	167	519
^	State3	138	618	1803

		State Determined by Viterbi		
		State1	State2	State3
Actual State	State1	1328	1455	283
^	State2	12	129	648
^	State3	29	51	2479

Table 2. Confusion matrices for Viterbi labelling of the 3 states - a) using single Gaussian parameter distributions; and b) using double Gaussian parameter distributions.

The experiment shows that Hidden Markov modelling is capable of reliably segmenting speech material into syllables but that discrimination between stressed and unstressed syllables needs to take into account further information, specifically syllable duration.

5. REFERENCE

Wagner, M., McKay, B., Sampath, S. & Slater, D. (1990) *Modelling the Suprasegmental acoustic Parameters of Some Phrase Structures of English*, Proc. Europ. Sig. Proc. Conf., EUSIPCO-90, Barcelona.