AUTOMATIC SEGMENTATION OF SPEECH SIGNALS INTO ARABIC SYLLABLES

A. Al-Otaibi* and Y. El-Imam+

* Electronics Department, Engineering Division
Kuwait Institute for Scientific Research

+ Kuwait Scientific Center
International Business Machines

ABSTRACT - Due to the nature of the Arabic language in terms of the rules governing formation of syllables by phonemes, and the one-to-one correspondence between the acoustics and the phonetics of Arabic syllables, a syllabic based approach for speech recognition of the Arabic language has a high potential for success. An experimental evaluation of an automatic speech segmentation algorithm into Arabic syllabic units is reported here. The parameter used for segmentation is the energy of the acoustic signal. Speech data consisting of mono-syllabic and multi-syllabic words were used to test the automatic Arabic syllabic segmentation algorithm. The algorithm has the advantage of being simple to implement.

INTRODUCTION

Speech is a complex acoustic signal (phonetic plus prosodic information mixed together) that results from the interaction of two independent functions, namely, the vocal tract system function and the excitation function. It is characterized by being of a continuant nature and of high data rate. It is therefore very advantageous to segment the steady stream of speech samples into minimal segments corresponding to minimal sound units (phonemes) in order to ease the computation of the characteristic speech parameters and to be able to group together acoustically homogenous segments. Because of speech variability and other problems such as noise, one to one correspondence between the acoustic events, represented by the speech signal and the corresponding phonemes, does not exist. It is therefore extremely difficult to automatically segment speech into phonemes by locating the beginning and end locations of each phoneme present in the utterance. In practice, two methods of speech segmentation have been used, namely, segmentation into constant time frame intervals (e.g., 10 msec. duration), and segmentation into one phonetic class segments based on extraction of certain acoustic parameters (e.g., voiced/unvoiced flag).

Syllables are important sound units for speech recognition, because coarticulation effects are already included in these units and also because they represent perceptual units.

Syllables may be defined linguistically in terms of the inherent sonarity of each sound (Ladefoged, 1982). Peaks of sonarity generally correspond to peaks of syllabicity (syllable nucleus), but one cannot empirically extract sonarity information alone from the acoustic signal as other information exist in parallel such as prosody (stress and pitch). This mix-up of prosodic information and sonarity makes segmentation into syllables for the English language a rather difficult task to accomplish accurately. Arabic, on the other hand, has well defined rules governing grouping of phonemes into syllables, formation of syllables within a word and placement of stress on individual syllables within a word. It is thus inclined more towards simple decomposition into syllabic units.

The high degree of correspondence between the phonetics and the acoustics of the Arabic language can be observed in Figure 1 for the Arabic word /jataʕallam/. This figure shows the pressure waveforms of /jataʕallam/ and its respective energy contour in the energy profile. This utterance is composed of four syllables, two of the type CV, and two of the type CVC, as indicated in Figure 1. Since the consonants associated with these syllables are now of the type semi-vowel /j/, stop /t/, pharyngeal /ʔ/, liquid /l/ and nasal /m/, respectively, the energy contour exhibits many maxima and minima points. If we examine closely, the energy contour and pick up the most prominent maxima (highest peak among several peaks that are close together), and if these prominent maxima are far enough form each other (›80 msec.), we can then assume that the syllable count for this utterance is equal to the number of the prominent energy maxima points.

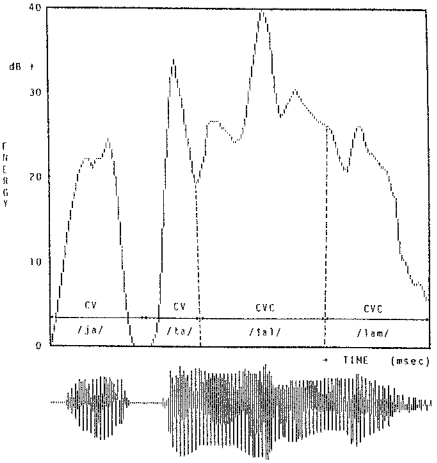

Figure 1. Pressure waveform and energy contour of the Arabic word /jataʕallam/.

SEGMENTATION ALGORITHM

The segmentation algorithm proceeds by performing four steps, namely, digital-filtering the speech samples, computing the speech utterance energy profile, algorithm for maxima/minima points-picking from the energy profile and maxima/minima points significance testing. Each step is implemented by a software procedure. The following is a brief description of each subroutine procedure:

- Filtering

  In order to smooth the energy contour of the speech signals, the speech samples time signal is low-pass filtered by a recursive digital filter. Since prominent energy peaks correspond to Arabic vowel regions and since also the first formant F1 is always the strongest among the three vocally generated formants F1, F2 and F3 (Carter, 1984), the 3dB cutoff frequency of the digital filter is set at 1900 Hz.

- Speech Energy Profile

  The energy E(k) of the speech signals is calculated in frames of 256 speech samples in length, windowed by a Hamming window w(n) of size 256, and overlapped by a quarter frame length.

- Maxima/Minima Points Picking Algorithm

  This procedure is responsible for picking all the maxima and minima points of the energy contour. A maxima energy point is determined once the energy slope sign changes from positive to negative and vice versa for the minima energy point.

  Generally, there will be many more maxima than the number of syllables. This can be attributed to the nature of speech signals in terms of its complexity since, besides phonology, it carries other information such as prosody and physiology. Maxima and minima of the energy contour are potential syllable nuclei and syllable boundaries, respectively.

- Maxima/Minima Points Significance Testing

  In order to arrive at potential syllabic nuclei and boundaries, some form of significance testing has to be applied to the collected maxima and minima points. This is equivalent to smoothing the energy curve by eliminating false and misleading maxima and minima points created by the complex nature of the speech signal.

  In order for the syllabic segmentation process to be effective, the evaluation measures have to be context dependent. Basically, the collected maxima and minima points of the energy contour are subjected to three kinds of measures. Each measure can be looked at as one level of a smoothing process. The measures are:

1. In a set of up to three consecutive maxima points within 19.2 msec., the highest energy maxima point is selected and the rest is rejected if the pairs of maxima and minima points form a positive slope. Similarly, in a set of up to three consecutive minima points, the lowest energy minima point is selected and the rest are rejected if the pairs of maxima and minima points form a negative slope.

2. Two threshold constants, dx and dy are calculated. dx=4% of total utterance duration and dy=10% of the maximum energy frame. These threshold consonants are used to evaluate the significance of each maximum point and its associated minima points by detecting if they fall inside these thresholds.

3. Minima points of energy values above 70% of the maximum energy frame are rejected. Similarly, maxima points of energy values less than 15% of the maximum energy frame are rejected.

TEST SPEECH VOCABULARY

The speech utterances used to test the Arabic syllabic segmentation algorithm consisted of 50 monosyllabic and 82 multisyllabic words spoken by three males and one female, referred to as informants. Each informant made on the average five repetitions per word. The uttered words are digitized via a developed PC-based Arabic speech processing system (Al-Otaibi, 1986). Recordings of uttered Arabic words were not made in one session, hence the effect of time variations is included.

SYLLABIC SEGMENTATION RESULTS

The performance of the Arabic syllabic segmentation algorithm described above was evaluated by processing the 50 monosyllabic and 82 multisyllabic words, spoken by the informants. Words chosen were those having the variety of syllable types and syllable structures of the Arabic language. The testing process of the Arabic syllabic segmentation algorithm proceeded in the following manner:

1. First the digitized Arabic words were used to tune-in the threshold constraints values. The following threshold constraints values were found to produce satisfactory Arabic syllabic counts.

   - dx = 4% of total utterance duration.
   - dy = 10% of the maximum energy frame recorded, (i.e., 10% of maximum maxima point energy).
   - The upper limit energy value of any minima point to be considered as potential syllable boundary is equal to 70% of the maximum energy frame.
   - The lower limit energy value of any maxima point to be considered as potential syllable nucleus is equal to 15% of the maximum energy frame.

The dx threshold value was found to be more critical than the others in obtaining correct syllabic count because it relates to syllable duration, i.e., low dx values result in extra syllabic counts, i.e., syllable fragments are generated. On the other hand, the other threshold values were found to be less critical in affecting syllabic counts. For example, the dy value can vary between 8-12% of the maximum energy frame. Similarly, the upper and lower energy limits of the minima and maxima points may vary between 65-70% of maximum energy frame, and 10-15% of maximum energy frame, respectively.

2. Secondly, the performance of the Arabic syllabic segmentation algorithm was tested. The accuracy of Arabic syllabic counting using the 50 monosyllabic words was 96%, using 39 di-syllabic words 84%, and using 43 tri-syllabic words 93%. These results are shown in the form of a confusion matrix displayed in Table 1.

   Extra syllable counts are caused by syllabic structures containing the consonant /r/, because the phoneme /r/ possesses distinct formant structures which are interrupted by a short gap. The gap represents a cut-off of energy as the tip of the tongue taps against the hard palate. The energy contour will therefore have a minima point which is picked up by the maxima/minima points picking algorithm.

The overall accuracy of the segmentation process is 92%. The syllabic boundaries located by the algorithm are simply related to the actual syllabic boundaries.

CONCLUSION

An automatic segmentation algorithm suitable for partitioning Arabic speech into Arabic syllabic units has been developed. The algorithm was tested with monosyllabic and multisyllabic Arabic words and its overall performance was proved to be satisfactory. The main parameter affecting the reliability of the derived syllabic count are the threshold values dx and dy, which are context sensitive. The segmentation algorithm is computationally simple to realize and execute, as will be demonstrated later. It has many important applications in Arabic speech processing, especially in speech recognition.

REFERENCES

Al-Otaibi, A. (1986) *PC-Based Arabic Speech Processing System*, First National Conference on 'Computers and its Applications in Jordan', November 2-5, Amman, Jordan.

Carter, J.P. (1984) *Electronically Hearing: Computer Speech Recognition*, Howard W. Sam & Co. Inc., Indiana, USA.

Ladefoged, P. (1982) *A Course in Phonetics*, Harcourt Brace Jovanovich, Inc., United Kingdom.

SYLLABIC COUNT

Actual Syllable	1	2	3	4	Count Accuracy
1	48	2	-	-	96%
2	1	33	5	-	84%
3	-	2	40	1	93%

Overall Accuracy = 92%

Table 1. Syllabic Count Confusion Matrix.