Proceedings of The Institute of Acoustics

ALIGNING SPEECH WITH TEXT

D.R. Miller and S.D. Isard

Lab. of Experimental Psychology, University of Sussex

The system described here aligns the speech wave of an utterance with a phonetic transcription. The input to the system is a digitised waveform, together with a transcription, and the output is a set of pointers into the waveform giving the locations of the phones mentioned in the transcription.

We have in mind three applications for such a system:-

a) Construction of diphone dictionaries for synthesis of speech on a variety of voices and dialects.

b) As a verification module in a speech recognition system.

c) As an aid in manipulating utterances to construct stimuli for laboratory experiments.

The system proceeds through four stages:-

1) Extraction of acoustic parameters for use by the later stages of the system.

2) Segmentation of the wave into sub-phonetic units, separated by "acoustic edges". A phone should consist of one or more of these units

3) Grouping these sub-phonetic units into phone-size units using a system of rules which take local context into account.

4) The use of these units to form a lattice of syllable nypotheses, which can then be matched against the phonetic transcription.

PARAMETER EXTRACTION

The parameters are extracted for two purposes: first they are used for locating acoustic edges (stage 2) and then for establishing the phonetic nature of the material between the edges (stage 3).

Our parameters are:-

1. First formant frequency (F1)

2. Second formant frequency (F2)
3. Third formant frequency (F3)
4. A measure of the overall signal energy.
5. A voicing indicator.
6. The energy in the 0 - 1Khz band (Low Freq Energy)
7. The energy in the 1 - 2.5Khz band (Mid Freq Energy)
8. The energy in the 2.5 - 5Khz band (High Freq Energy)

A 5Khz low pass filter is applied to the speech wave, which is then sampled at a rate of 10Khz. F1, F2 and F3 are extracted using root solving techniques on the result of a 12 pole linear predictive analysis.

The voicing indicator is a simple zero-crossing counter (ZCR) which, although inadequate for any fine voicing decisions, gives a gross indication for each 10msec window.

As a measure of overall signal energy the average magnitude function (AMF) is used

$$AMF = \sum_{m=1}^{n} |X(m)|/n$$

where X(m) = the m^{th} sampled data point.

Because we are interested in the underlying trend of these parameters rather than the fine local detail the computed values of these parameters are smoothed. The smoothing operation consists of a five point running median followed by a three point running median followed by a Hann window. This is similar to the method described by Tukey [1] and recently reported on by Gallagher and Wise [2].

SEGMENTATION

The segmentation is pre-categorical in that the 10 msec frames of speech are grouped into larger units on the basis of acoustic stability, without any attempt to attach phonetic labels to the frames themselves. The idea underlying this approach is that a movement of the articulators from the target position associated with one phone to that of the next will be correlated with a peak in the rate of change of the acoustic parameters. We refer to these peaks as acoustic edges.

The algorithm used for the segmentation is a modification of a stability function proposed by Lienard et al [3], in which various

speech parameters are compared against themselves over time; (a simular approach has also been used by Neel et al [4]).

We use an 'instability' function I(t):-

$$I(t) = \frac{\sum |P(m,t-\Theta) - P(m,t+\Theta)|}{\sum P(m,t-\Theta) + P(m,t+\Theta)}$$

where

P(m,t) is the value of the m^{th} parameter at time t, and
Θ is a time shift.

In principle Θ could be chosen for each parameter individually, but on the basis of trials with Θ values in the range 10 - 100 msec in 10 msec steps we have settled on a uniform value of 20 msec for all parameters.

Acoustic edges are points at which I(t) achieves a maximum and segments run from one acoustic edge to the next. The stable point of a segment is the point between its edges where I(t) reaches a minimum.

Once boundaries and stable points have been established, we compute for each segment:-

1. The value of F1, F2, F3, ZCR, and AMF at the stable point of that segment.

2. A measure of the change of each parameter across the segment, defined as:-

$$\frac{P(b)-P(a)}{P(b)+P(a)}$$

where P is the parameter in question and a and b are the endpoints of the segment.

Acoustic feature labels are assigned on the basis of these values. The labels stand for combinations of conditions on the values. For instance the label FRIC summarizes the conditions:-

The energy in the high frequency band exceeds the energy in the low and mid freq bands and the zero crossing rate is high.

We have given these features mnemonic names related to phonetic catagories for which they often serve as cues. However the assignment of the label FRIC to a segment at this stage does not necessarily mean that the segment will be proposed as part of a fricative phone later on.

ALIGNING SPEECH WITH TEXT

PHONETIC LABELLING AND GROUPING

Our goal here is to group the acoustic segments into phone-sized units labelled by phonetic features. Whereas at the previous stage segments where located first and labelled afterwards, here labelling and grouping are a single process. Groupings are proposed with an eye toward what type of phone they might turn out to be. Each grouping rule is associated with a phonetic feature and says in effect that a given group of segments is potentially a phone bearing that feature label.

The feature labels used are:-

1. VOWEL
2. VOICED STOP
3. VOICELESS STOP
4. NASAL
5. FRICATIVE
6. GLIDE

Stops and nasals may be further specified as pre-vocalic or post-vocalic. While this set of features does not fully identify the phones, it is sufficient for grouping them into syllables and for anchoring the match of the syllables with the phonetic transcription. It should be noted that the label Glide is used for all consonants that are not members of the other categories.

The grouping rules are applied first to potential vowels and then to form initial and final consonant clusters around them. Identification of vowels is done just on the basis of the acoustic features assigned to their segments. Consonants are formed taking into account both their own acoustic features and the assignments that have been made to the material between them and the vowel. An example of such a rule is:-

> If this segment has the acoustic label WEAK NASAL and the prior segment had the acoustic label CODA NASAL and the segment before that forms part of a group with the phonetic label VOWEL then this segment and the last segment are grouped and given the phonetic label POST-VOCALIC NASAL.

Grouping rules for consonants adjacent to the vowel have been completed; some of those dealing with segments further out are still under development.

Sometimes more than one candidate labelling is produced. We are more concerned with making certain that the correct answer occurs

among the candidates than with avoiding false alarms, since the ultimate match against the phonetic transcription will seek out the phones that are supposed to be present and ignore the rest.

SYLLABLE FORMATION

Only some sequences of phones constitute possible syllables. The sequence <fricative nasal vowel stop> could be a syllable but the sequence <nasal fricative vowel stop> could not. Our algorithm finds the segments that have been labeled as vowels, then searches around them to form all allowable consonant clusters.

All the possible syllables which can be formed from the labelled segments are returned.

The input transcription is also grouped into syllables and used as the pattern to match against the syllable hypotheses found by the analysis. There is an optional (and fallible) preproccesor for turning English text into a phonetic transcription.

CONCLUSION

The system has been tested on a corpus of about 100 utterances constructed to contain just those combinations of phones that the grouping rules have so far been written for. The corpus has been spoken by three male workers in our laboratory, and the system's segmentation has been in good agreement with that done independently by a phonetician.

Ref:-

[1] T.W. Tukey, 'Nonlinear (nonsuperposable) methods for smoothing data', Congress Record, EASCON, 1974.

[2] N.C. Gallagher and G.L. Wise, 'A theoretical analysis of the properties of median filters', IEEE Trans ASSP, vol ASSP-29, no 6, 1981.

[3] J.S. Lienard, M. Mlouka, J.J. Mariani, J. Sapaly, 'Real time segmentation of speech', in G. Fant (ed), Speech Communication Vol 3, Stockholm: Almqvist and Wiksell 1974.

[4] F. Neel, M. Eskenazi and J.J Mariani, 'A method to automatically constitute phonetic dictionaries for speech comprehension system', 105th meeting: Acoustical Society of America, J. Acoust. Soc. Am. Suppl. 1, Vol 73, Spring 1983, pp. S88.