Automatic labeling system using speaker-dependent phonetic unit references

AUTOMATIC LABELING SYSTEM USING SPEAKER-DEPENDENT PHONETIC UNIT REFERENCES

SHOZO MAKINO
HISASHI WAKITA

Speech Technology Laboratory
3888 State Street, Santa Barbara, CA. 93105, USA

ABSTRACT

This paper describes a new automatic labeling system using speaker-dependent reference patterns for 73 phonetic units in American English. The system segments arbitrary utterances into phonetic units and automatically adapts to a new speaker using a small set of training words. The labeling of the training words begins with the words which can be easily segmented into necessary phonetic units and then reference patterns for each unit are computed by use of vector quantization clustering. Using the training reference patterns together with vocalic-consonant information, the speech input is aligned with the transcription using dynamic programming with duration constraints for each phonetic unit. More accurate phonetic boundaries are obtained using new reference patterns derived from the input speech.

The system was evaluated on 15 repetitions of 104 words uttered by two males and one female. Standard deviation of differences between manually labeled and automatically obtained boundaries ranged from 21 ms to 27 ms. Most of the discrepancies occurred at the boundaries between vowels, nasals and liquids.

INTRODUCTION

A large amount of segmented and labeled speech data with the corresponding phonetic transcriptions are essential for developing a speech recognition system based on phonetic units, as well as for developing a text-to-speech synthesis system. Traditional manual labeling is extremely time consuming and subject to lack of consistency and reproducibility of the results.

Over the past few years, several automatic time alignment procedures have been proposed in the literature. Most of these approaches attempt to align the input utterance with a manually labeled reference utterance using dynamic time-warping[1,2]. A second approach, which also uses dynamic programming, is to segment and label the utterance into broad phonetic classes independent of speakers prior to time alignment[3,4]. Although broad phonetic classes are robust, much better labeling results are expected if a system is speaker-dependent and the reference patterns for phonetic units can be made from input speech.

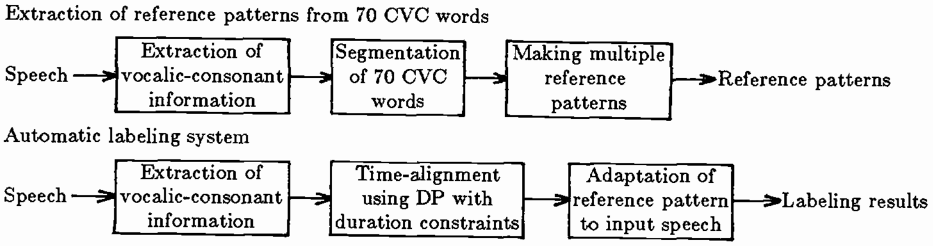
Extraction of reference patterns from 70 CVC words
Speech
Extraction of vocalic-consonant information
Segmentation of 70 CVC words
Making multiple reference patterns
Reference patterns
Automatic labeling system
Speech
Extraction of vocalic-consonant information
Time-alignment using DP with duration constraints
Adaptation of reference pattern to input speech
Labeling results

Fig. 1 Schematic diagram of automatic labeling system

This paper describes an approach to such a labeling system which segments arbitrary word utterances into phonetic units and automatically adapts to a new speaker using a small set of training words. Using the training reference patterns and vocalic-consonant information, the input speech is first aligned with the transcription using dynamic programming with duration constraints for each phonetic unit, and then more accurate phonetic boundaries are determined using new reference patterns derived from the input speech.

OUTLINE OF THE AUTOMATIC LABELING SYSTEM

Figure 1 shows a schematic diagram of the automatic labeling system. The speech signal is digitized at 10 kHz. Logarithmic power, zero crossing rate and 10 LPC cepstral coefficients are computed once every 10 ms.

Reference patterns are made from 70 CVC training words which contain most word-initial and word-final consonants and stressed vowels. The end-point of the speech is determined using the logarithmic power and the zero crossing rate. Vocalic-consonant information is extracted using Laplacian filter outputs of the logarithmic power, the zero crossing rate and the first cepstral coefficient. Using the frame with maximum power and the consonant segments detected by the vocalic-consonant information, the labeling of the 70 CVC words begins with the words which can be easily segmented into necessary phonetic units. In the process of labeling, the reference patterns of phonetic units are temporarily generated from the training speech and used for obtaining boundaries. After the labeling of the 70 CVC words, reference patterns for each unit are computed using vector quantization clustering.

From the input speech to be labeled, end-points of the speech and the vocalic-consonant information are extracted in a similar fashion, where the end-point detection algorithm generates at most two candidates for the beginning frame and for the end frame of an input speech. The endpoint is finally determined using these

CH2243-4/86/0000-2783 $1.00 © 1986 IEEE

candidates and continuous DP[6]. Using only the training reference patterns of units corresponding to the given transcription for an input speech, the input is aligned with the given transcription using dynamic programming with duration constraints for each phonetic unit[7], where the distance between a phonetic unit and an input frame is weighted based on vocalic-consonant information. The final phonetic boundaries are determined using new reference patterns derived from the input speech by applying the dynamic programming with duration constraints.

EXTRACTION OF REFERENCE PATTERNS

73 phonetic units derived from 50 phonemes are used for the labeling. 48 of the 50 phonemes are included in ARPA-BET. Two other phonemes are unstressed vowels. Each voiceless stop and voiced stop in the 50 phonemes has three phonetic units, corresponding to buzz portion, burst portion and aspiration portion. Each diphthong has two phonetic units, corresponding to the first half portion and the remaining portion. 70 CVC words contain most word-initial and word-final consonants and stressed vowels. The remaining phonetic units which cannot be extracted from the 70 CVC words are substituted by similar phonetic units. An end point is detected using the logarithmic power and the zero crossing rate. Vocalic-consonant parameters are computed for each frame using a Laplacian filter as follows:

$$y_i(j) = \sum_{k=-7}^{7} w(k)\, x_i(j+k), \quad \text{for } i = 1,2,3.$$

$$\text{where } x_1(j) = 10.0 * \frac{pow(j) - \min\limits_{1<k<J} pow(k)}{\max\limits_{1<l<J} pow(l) - \min\limits_{1<k<J} pow(k)}$$

$pow(j)$: Logarithmic power of the j-th frame
$x_2(j)$: Zero crossing rate of the j-th frame
$x_3(j)$: The first LPC cepstral coefficient of the j-th frame
$w(k)$: Weighting coefficient from Laplacian filter
J : Length of input speech

Vocalic-consonant decision for the j-th frame is made as follows:

$$Voc(j) = \text{'C'}, \text{ if } y_1(j) < -5.0$$

$$Voc(j) = \text{'U'}, \begin{cases} \text{if } y_2(j) > 350.0, \text{ or} \\ \text{if } y_3(j) < -5.5, \text{ or} \\ \text{if } x_1(j) < 2.0, \text{ or} \\ \text{if } x_2(j) > 80.0 \end{cases}$$

The frame without 'C' or 'U' is regarded as a vocalic frame. If the frame is assigned to both 'C' and 'U', the frame is regarded as 'U'. Going back from the frame with maximum power to the beginning frame of the speech, the first frame with 'C' or 'U' is regarded as the final frame of the word-initial consonant segment. If the frame with 'C' or 'U' is not found, the middle frame between the beginning frame of the speech and the frame with maximum power is regarded as the beginning frame of the vowel segment. Going forward from the frame with maximum power to the end frame of the speech, the first frame with 'C' or 'U' is regarded as the beginning frame of the word-final consonant segment. If the frame with 'C' or 'U' is not found, the middle frame between the frame with maximum power and the end frame of the speech is regarded as the final frame of the vowel segment. If the vowel is not a diphthong and is not preceded by liquids, the frame with maximum power is regarded as a typical frame of the vowel segment. If the vowel is not a diphthong and is preceded by liquids, the eighth frame after the frame with maximum power is regarded as the typical frame. If the vowel is a diphthong, the frame with the nearest distance in the vowel segment from a similar single vowel is regarded as the typical frame. Accordingly the two typical frames are extracted from the vowel segment for a diphthong. A typical frame for the word-initial

Table 1 Parameters for extraction of typical frames and boundaries in consonants

Phoneme group	Typical frame		Boundary	
	Word-initial	Word-final	Word-initial	Word-final
y,w,r,l	f_1	f_3	f_4	f_4
m,n,nx	f_1	f_2	f_4	f_4
p,t,k,ch,jh	$\max \frac{dC_0}{dt}$	$\max \frac{dC_0}{dt}$	f_4	f_4
b,d,g,dx	$\max \frac{dC_0}{dt}$	$\min \frac{dC_2}{dt}$	$\max \frac{dC_1}{dt}$	$\max \frac{dC_2}{dt}$
hh,f,th,s,sh, v,dh,z,zh	$\min C_1$	$\min C_1$	f_4	f_4

f_1: The 5 frames after the beginning frame of the speech
f_2: The 5 frames before the end frame of the speech
f_3: The frame with the nearest distance from the reference
f_4: The frame with equal distances from the typical frame of the vowel and the typical frame of the consonant

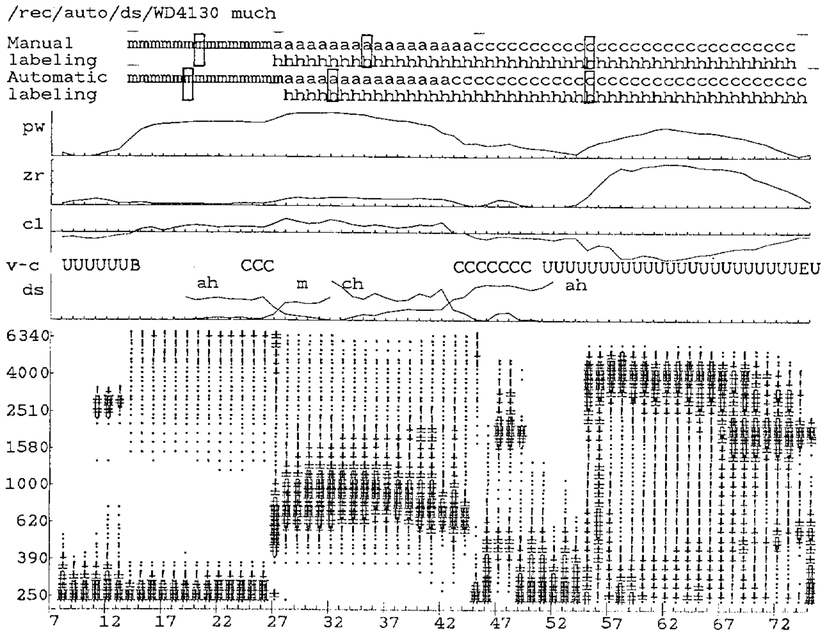

Fig. 2 An example of labeling for /m ah ch/

51. 9. 2

consonant is extracted from the segment between the beginning frame of the speech and the typical frame of the vowel segment using the parameter as shown in the table 1. A typical frame for the word-final consonant is extracted in a similar fashion. The labeling of the 70 CVC words begins with the words which can be easily segmented into necessary phonetic units. In the process of labeling, the reference patterns of phonetic units are temporarily generated and used for obtaining typical frames. Figure 2 shows an example of the labeling for /m ah ch/. 'B' and 'E' indicate the beginning and end frames of the speech. The frame with maximum power is the 32-nd frame. The typical frame for /m/ is defined as the 5-th frame after the beginning frame of the speech, that is, the 19-th frame. The frame(28-th) with equal distances from both typical frames is regarded as the boundary. The frame(55-th) with maximum derivative of the logarithmic power(C_0) is regarded as the burst frame of /ch/. The typical frame for the buzz portion of /ch/ is defined as the 3-rd frame before the burst frame. The frame(43-rd) with equal distances from the typical frames of the vowel and the buzz portion is regarded as the boundary.

After the labeling of the 70 CVC words, reference patterns for each unit are computed by means of vector quantization clustering[5] from the speech samples which are gathered up the 5 frames around the typical frame of the corresponding phonetic unit. At most 8 reference patterns per unit are extracted, where a reference pattern is represented by 10 cepstral coefficients.

AUTOMATIC LABELING USING VOCALIC-CONSONANT INFORMATION AND DYNAMIC PROGRAMMING WITH DURATION CONSTRAINTS

The end-point detection algorithm generates at most two candidates for the beginning frame and for the end frame of the speech. The end-point is finally determined using continuous DP[6].

The input speech whose end-points have been detected is then time-aligned with its phonetic description using dynamic programming with duration constraints[7]. A reference template is made by concatenation of the phonetic units with its maximum duration. A flag sequence is similarly made by '0' or '1' with the maximum duration of the corresponding phonetic unit as shown in Fig.3, where the first portion of the minimum duration in each sub-sequence corresponding to the phonetic unit has '1'. The remaining portion of the sub-sequence has '0'. The duration-constraints DP is calculated as follows:

1. Initialize
$D(1,j) = \infty$ for $j = jb+1, jb+2, \dots, je$
$D(1,jb) = d(1,jb)$

2. Execute 3 for $i = 2,3,\dots,I_d$
3. For $j = jb+1, jb+2, \dots, je$.

$$D(i,j) = \min\begin{cases} D(i-1,j-1) + d(i,j) & \text{if } flag(i) = \text{'1' or '0'} \\ D(i-1,j) & \text{if } flag(i) = \text{'0'} \end{cases}$$

```
Phonetic     fffffffffiiiiiiiiiiiiiiiidddddddddd
units                 yyyyyyyyyyyyyyyy
Flag         11110000011111110000000001100110000
                                      buzz|aspiration
                                          burst
```

Fig. 3 An example of reference template for /f iy d/

jb : The beginning frame of the speech
je : The end frame of the speech
I_d : The length of reference template
$flag(i)$: The flag value for the i-th frame
$d(i,j)$: The nearest cepstral distance between the j-th frame of input utterance and the phonetic unit corresponding to the i-th frame
where $d(i,j) = ds*A$

$$ds = \min_{1 \le m \le M} ds_m$$

$$ds_m = \sum_{p=1}^{10} (C_{jp} - C_{mp})^2$$

C_{mp} : The m-th reference pattern of the phonetic unit corresponding to the i-th frame

The distance between the i-th reference frame and the j-th input frame is defined as the weighted distance which is computed by the distance(ds) with weighting coefficient(A). The distance(ds) is defined as the minimum distance in the M distances to the j-th input frame computed by M-reference patterns for the phonetic unit corresponding to the i-th reference frame. The weighting coefficient(A) is dependent on vocalic-consonant information of the j-th input speech frame and the phonetic unit of the i-th reference frame. The beginning and the final frames of each segment are defined by going back from the end frame (je) along the optimal path.

Using the phonetic segment information obtained from the duration-constraints DP, the reference pattern of the phonetic unit is renewed. The three-frame averaged distances from the corresponding phonetic unit are computed for each frame in the segment. The three frame averaged cepstral coefficients around the frame with the minimum distance are regarded as the new reference pattern of the phonetic unit. The distances from the new reference pattern are recomputed from the beginning frame of the preceding segment to the final frame of the following segment. Based on the recomputed distances, the duration-constraints DP is applied again. The beginning and the final frames of each segment are finally defined by going back from the end frame(je) along the optimal path.

LABELING EXPERIMENTS

Five methods of automatic labeling were evaluated on 15 repetitions of 104 words uttered by two males and one female.

In method 1, input speech is time-aligned with the corresponding word reference template in the first repetition uttered by the same speaker using a dynamic time warping method[8]. In method 2, the 73 phonetic units and duration constraints are used for the labeling without vocalic-consonant information. In methods 3 through 5, vocalic-consonant information is used for the automatic labeling. In addition, methods 4 and 5 employs adaptation to the input speech, that is, new reference patterns for the phonetic units are derived from the input speech using the results obtained by method 3. Methods 1 through 4 use the manually extracted end-points, whereas method 5 uses the automatically extracted end-points.

Four types of reference patterns were also tested. The first type (Ref 1) is a single pattern per phonetic unit, manually extracted from the 104 keyboard vocabulary. The other types are multiple patterns per phonetic unit and are computed by means of vector quantization clustering. The second type(Ref 2) is computed from speech samples extracted from manually labeled typical frames in the 104 keyboard vocabulary. The third (Ref 3) is computed from speech samples extracted from manually labeled typical frames in the 70 CVC training words. The fourth (Ref 4) is computed from speech samples which are derived from automatically extracted typical frames in the 70 CVC training words.

Table 2 shows the comparison between the labeling methods. The values in the table represents the averaged difference between the method's standard deviations. The standard deviations of the distributions of differences between manually and automatically labeled boundaries for 104 words were averaged over the 15 repetitions and the difference taken. The comparison between methods 2 and 3 shows that the vocalic-consonant information remarkably improves the accuracy. The comparison between methods 3 and 4 shows that the adaptation to the input speech also improves the accuracy. Method 4 which gave the best results is almost equal to method 1 based on word reference patterns. However, method 5 indicated rather poor results due to insufficient end-points detection.

Table 3 shows the comparison between reference patterns. Use of multiple reference patterns per phonetic unit is very effective. Ref 3 gave nearly the same results as those obtained by Ref 2 except the female voice. This indicates that the 70 CVC words contain sufficient information for making multiple reference patterns per unit. The large difference for the female is due to the fact that the 104 key-board vocabulary was recorded 3 years earlier than the 70 CVC words, and, in addition, the subject had a slight cold when the 70 CVC words were recorded. Automatic extraction of typical frame from the 70 CVC words works well as indicated by the fact that the results obtained by Ref 4 were nearly the same as those obtained by Ref 3.

Standard deviation of differences between manually labeled and automatically obtained boundaries ranged from 21 ms to 27 ms. Most of the discrepancies occurred at the boundaries among vowels, nasals and liquids. Methods 4 and 5 with Ref 4 were applied to another vocabulary with 275 words uttered by the first male speaker. The deviations are 24 ms and 27 ms for methods 4 and 5. These values are almost the same as those for the 104 key-board vocabulary.

Table 2 Comparison between labeling methods(unit: ms)

Method	Type of reference			
	Ref 1	Ref 2	Ref 3	Ref 4
Method 2 - Method 3	2.6	1.7	3.2	3.9
Method 3 - Method 4	4.6	1.3	1.5	2.3
Method 4 - Method 5		-2.0	-1.3	-1.5
Method 1 - Method 4	-3.5	-0.2	-1.4	-0.5
Method 1 - Method 5		-2.7	-3.3	-2.5

Table 3 Comparison between reference patterns(unit: ms)

Reference	Sub.	Type of method			
		Method 2	Method 3	Method 4	Method 5
Ref 1 - Ref 2	M 1	5.2	4.2	1.6	
Ref 2 - Ref 3	M 1	-2.8	-1.2	-1.0	-1.0
	M 2	-1.6	0.6	0.0	0.0
	F 1	-10.8	-5.6	-4.8	-3.0
	Avr.	-5.1	-2.1	-1.9	-1.3
Ref 3 - Ref 4	M 1	-0.8	-1.6	-0.4	0.0
	M 2	-1.6	-0.8	0.6	0.0
	F 1	0.4	-0.2	0.4	0.2
	Avr.	-0.5	-0.9	0.2	0.1

CONCLUSION

The system gave the promising results. However the end-point detection should be improved and temporal features should be introduced into the system in order to improve the accuracy for liquids and nasals.

REFERENCE

[1] R.M. Chamberlain and J.S.Bridle, "Zip; A dynamic algorithm for time-aligning two indefinitely long sequences." in Proc. ICASSP, Apr.1983, pp.816-819.

[2] H.D. Hohne, C. Coker, S. E. Levinson and L. R. Rabiner,"On temporal alignment of sentences of natural and synthetic speech," IEEE Trans. Acoust., Speech, Signal Processing, vol.ASSP-31, pp.807-813, August 1983.

[3] M.Wagner, "Automatic labeling of continuous speech with a given phonetic transcription using dynamic programming algorithms,"in Proc. ICASSP, Apr. 1981, pp.1156-1159.

[4] H.G. Leung and V.W. Zue,"A procedure for automatic alignment of phonetic transcriptions with continuous speech,"in Proc. ICASSP, March 1984, pp.2.7.1-2.7.4.

[5] Y. Linde, A. Buzo and R.M. Gray,"An algorithm for vector quantizer design," IEEE Trans. Commun., vol.COM-28, pp.84-95, Jan. 1980.

[6] R. Oka,"An infinite-connected word recognition system for male speakers using time-space dynamic programming,"in Proceeding of the 5-th IJCAI, 1979.

[7] M. Kohda, S. Hashimoto and S. Saito, "Spoken digit mechanical recognition system," IECJ Trans., 55-D, pp.186-193, March 1972.

[8] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition" IEEE Trans. Acoust., Speech and Signal Processing, ASSP-22, pp.135-141, Feb. 1974.