IT ŽABLJAK '02

VII

naučno - stručni skup

INFORMACIONE TEHNOLOGIJE

SADAŠNJOST I BUDUĆNOST

Urednici

Novak D. Jauković
Srđan S. Stanković
Srbijanka R. Turajlić

INTERFEJS ZA PREPOZNAVANJE ODVOJENIH REČI NEZAVISNIH GOVORNIKA KOD IVR TELEFONSKIH SISTEMA

SPEAKER INDEPENDENT RECOGNITION OF ISOLATED WORDS INTERFACE FOR IVR TELEPHONE SYSTEMS

Ivan Kraljevski, Veterinary Institute Skopje, Skopje, R. Macedonia
Dragan Mihajlov, Faculty of Electrical Engineering – Skopje, R. Macedonia
Igor Stojanović, Customs Administration, Skopje, R. Macedonia

Sadržaj - *U ovom radu prestavljen je interfejs za prepoznavanje odvojenih reči nezavisnih govornika kod telefonskih sistema sa glasovnom interakcijom. Sistem za prepozanavanje govora je baziran na algoritmu za dinamičko vremensko zakrivljavanje - DTW radi veće računske efikasnosti. Ovaj sistem je implementiran i procenjen u simuliranim uslovima i rezultati pokazuju da je primenjenim metodama moguće postići razumne performanse.*

Abstract – *In this paper is given a proposal for a speaker independent speech recognition interface of isolated words with application in Interactive Voice Telephone systems. The overall structure of the recognition engine is based on the Dynamic Time Warping (DTW) paradigm for computational efficiency. The system is implemented and evaluated in simulated conditions and the results show that reasonable performance can be achieved by these methods.*

1. INTRODUCTION

Recent developments in speech technology have enabled a new generation of interactive voice response (IVR) services operating over the telephone network.

To properly identified and respond to user request, high performance speech recognition is required. Speech input achieved via the recognizer is usually supported with dial-pulse input. Speech output can be based on stored concatenated speech or speech to text system can be used. Speech recognition over telephony is a challenging task, since there is great signal, network and speaker variability.

In the following paper, we describe our work on developing speech recognition interface for general purpose (information services, banking, teleshoping etc.) telephone IVR system on Macedonian language.

2. SYSTEM OVERVIEW

The Fig. 1 presents an IVR system with an implementation of speaker independent speech recognition interface for isolated words.

2.1. Signal digitalization and preprocessing

The analog signal - telephone speech is limited to 3.3 kHz frequency range and although the transmission of speech through the telephone network degrades its quality significantly, it still remains largely intelligible. Thus 8 kHz sampling rate at 8 bits per sample corresponding to a bit-rate of 64 kbits is normally used.

After the process of quantization, the signal is normalized, noise reduction and automatic segmentation is performed with endpoint detection of spoken words. The method used to determine the endpoint uses the short time energy spectrum.

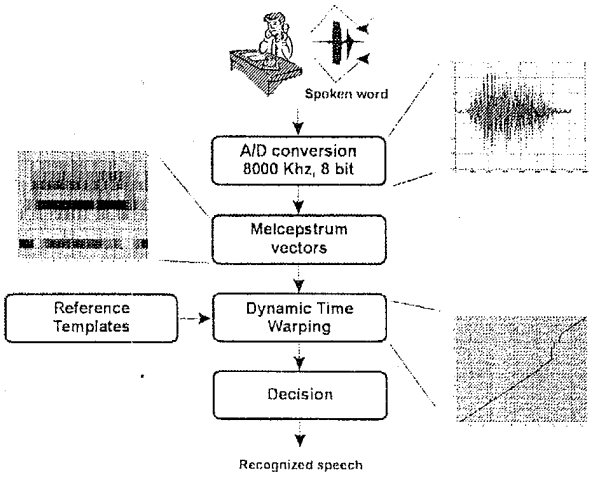


figure 1. Speech recognition IVR interface

The speech sequences used for evaluating system performance and estimating reference templates were obtained with manual segmentation of spoken utterances from various speakers and labeled by human expert (30% of them where used for reference templates and the others for performance evaluation).

2.2. Mel Frequency Cepstral Coefficients

In order to represent the spectral features of isolated words that has been segmented from the continuous signal stream, Fast Fourier Transformation (FFT) is computed for

each 16 ms with advance of 1/2 frame duration [1]. The standard Hamming window is applied to each frame. The signal is filtered with bank of triangular filters on Mel frequency scale (simulating the perception of the human hearing). Finally Discrete Cosine Transformation is applied and 12 MFCC coefficients for each frame were produced.

2.3. DTW – Dynamic Time Warping

After signal conversion from waveform to array of MFCC vectors in time domain, it is necessary to perform pattern matching algorithm to measure distance between input vector sequence and the reference templates. The word that refers to template with the smallest distance is the recognition result. Therefore, the recognition problem can be simplified as finding the distance between signal and template [2].

Since the feature vectors have multiple elements a means of calculating the local distance is required. The distance measure between feature vector x of signal 1 and feature vector y of signal 2 is given by the L2 or Euclidian distance metric:

$$d_2(\vec{x},\vec{y}) \equiv \sqrt{\sum_{k=1}^{p} |x_k - y_k|^2} \qquad (1)$$

Although the Euclidian metric is computationally more expensive than some metrics, it does give more weight to large differences in a single feature. It can also be shown that this metric has several desirable theoretical properties when comparing cepstra.

Speech is a time-dependent process. Several utterances of the same word are likely to have different durations, and utterances of the same word with the same duration will differ in the middle, due to different part of the word being spoken at different rates.

Classification of the input utterance is done by the criteria – simple global distance score between two vector arrays which is given by the sum of all local distances for the lowest distance path. The algorithm Dynamic Time Warping (DTW) always finds the path with minimal global distance – maximum likelihood between two utterances.

To reduce computational requirements and avoid evaluation of all possible paths which is extremely inefficient, as the number of possible paths is exponential in the length of the input. Some constraints are imposed to come up with efficient algorithm, instead, which is very important in real-time applications such as is IVR system:

- matching paths cannot go backwards in time;
- every frame in the input must be used in a matching path and
- global distance score is given by sum of local distance scores.

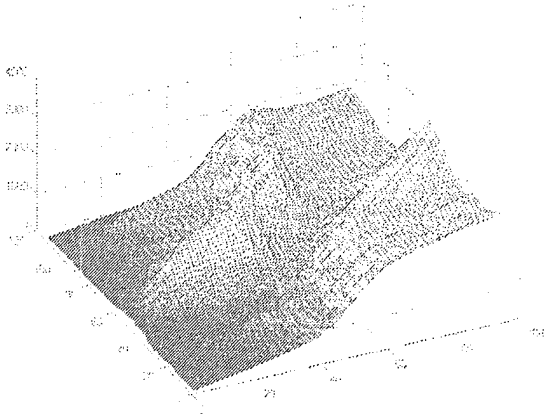
According to these constraints it is possible to compose recursive function for computing global distance score. If $D(i,j)$ is the global distance in (i,j) element of the time-time matrix and local distance in (i,j) is given by $d(i,j)$ then:

$$D\{1,1\} = d_2\{1,1\}, \qquad (2)$$

$$D\{i,j\} = \min\left[D\{i-1,j-1\}, D\{i-1,j\}, D\{i,j-1\}\right] + d_2\{i,j\}$$

value in (i,j) element is the minimal value from neighbor elements from the matrix added with the current local distance score between two vectors.

figure 2. The path to (i,j) can orginate form the three standard locations



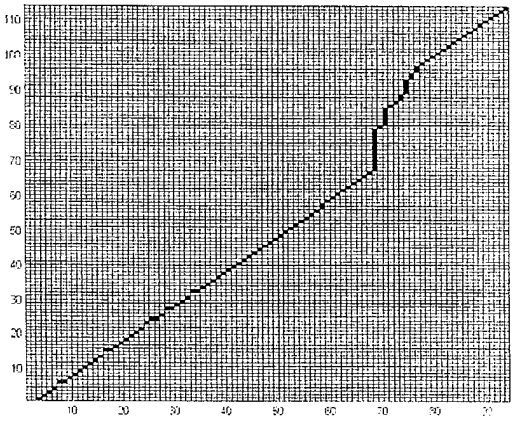


figure 3. Time-time matrix of D(i,j) function and the path of the time alignment of the utterances.

Basic requirement for an efficient recursive algorithm for computing $D(i,j)$ is the preset initial condition $D(1,1)=d(1,1)$. The final global distance $D(n,N)$ gives the overall matching score of the template with the input (n – length of the input vector sequence, N- length of template vector sequence). The input word is then recognized as the word corresponding to the template with the lowest matching score.

2.4. Clustering and template definitions

The pattern recognition approach to speech recognition requires no speech specific knowledge to be explicitly encoded into the system. This method is relatively ignorant of the choice of word, speaker, task or syntax. Therefore the reference templates are of great importance as these form the difference between a speaker dependent and a speaker independent system.

Recorded sequences were segmented, labeled and clustered manually by human expert in correspondent classes referring particular words, for example menu items of the menu driven IVR system.

It has been shown [3] that a few very carefully constructed templates can adequately represent a large speaker population for independent word recognition. There are two main approaches to this problem. The first approach is to estimate reference templates by simply averaging time aligned vector sequences - members of the same cluster.

Because of the great variability of the speech it is impossible to represent each word cluster with a single template. Second approach is to use multiple reference templates and compare input sequences with all templates of the all word clusters and select the cluster that gives the lowest possible global distance score, or by averaging all distance scores for each cluster, determines the one that provides lowest averaged score. The optimal number of multiple templates for each word can be estimated for given recognition rate using K-means algorithm [4].

3. SIMULATION RESULTS

For speech interface performance evaluation, 10 digits for 0 to 9 on Macedonian language were recorded with total of 190 spoken utterances from 8 different adult speakers (5 male and 3 female). These utterances were recorded in real environment with presence of ambient noise with sample rate of 22.05 KHz, 16 bit quality, and filtered with telephone band-pass filter (0.3-3.3 KHz) to simulate the behaviour of telephone channel.

In real conditions, channel cross talk and echoes might appear over the telephone link and signal processing techniques such as echo and adaptive noise suppression must be used to enhance the perceived signal. Noise reduction was performed for accurate endpoint detection. Then signal was downsampled to 8 KHz and coded with 8 bit quality.

The recognition is performed by comparing the input sequence with all multiple reference templates and the result is the word that refers to template with the lowest distance score. To make the system more robust and avoid cases where majority of the multiple reference templates refers to same word cluster, but the smallest distance score of the whole set refers to another cluster, additional ranking of results can be done and the cluster with highest number of matches is selected.

Two test types were carried out: speaker dependent and speaker independent. For speaker dependent test, each word cluster in the reference set contained 4 templates - instances of the same word of the same speaker. Testing was performed with 3 instances of the words that did not belong to reference set and the recognizer gives a result of recognition rate 100%.

For speaker independent test multiple reference templates were produced with one set of spoken digits from 4 different speakers each (2 male and 2 female). That gives a total of 40 reference templates. To estimate recognition rate of the recognizer, 120 utterances were divided in 3 groups:

1. sequences that are members of the reference set, the recognition rate is 100%;
2. known speakers, whose utterances were use for templates, recognition rate is 90%;
3. unknown speakers, where the system have to recognize the sequences only using available features from the reference set, recognition rate is 67.5%.

Table 1. Recognized words and the recognition rate for speaker independent system

	Words	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"	Match
Unknown speakers	Spk. 1 (M)	"0"	"1"	"0"	"4"	"4"	"5"	"6"	"7"	"8"	"5"	7
	Spk. 2 (M)	"0"	"5"	"0"	"4"	"4"	"5"	"6"	"7"	"8"	"5"	6
	Spk. 3 (F)	"0"	"1"	"2"	"3"	"4"	"5"	"5"	"1"	"8"	"9"	8
	Spk. 4 (M)	"6"	"7"	"0"	"3"	"4"	"5"	"6"	"7"	"8"	"5"	6
Word matching		75,00%	50,00%	25,00%	50,00%	100,00%	100,00%	100,00%	75,00%	100,00%	25,00%	67,50%
Known speakers	Spk. 5 (F)	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"	10
	Spk. 6 (M)	"0"	"1"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"5"	9
	Spk. 7 (F)	"0"	"4"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"5"	8
	Spk. 8 (M)	"0"	"7"	"2"	"3"	"4"	"5"	"6"	"7"	"8"	"9"	9
Word matching		100,00%	50,00%	100,00%	100,00%	100,00%	100,00%	75,00%	100,00%	100,00%	50,00%	90,00%
*F-Female, M-Male		✓	×	×	✓	✓	✓	✓	✓	✓	×	

Results given in Tab. 1 have shown that recognition rate for unknown speakers case was unacceptably low for application in real conditions. The reason for that is the speech signal type (low quality, noisy telephone speech signal) and inappropriate choice for reference templates. On the other hand, it must be noticed that some of the words were recognized with 100% recognition rate and others were recognizing as another word. The reason for that is similar sounding and they have similar spectral features of those words, as well as the order of phonemes of same type (plosives, vowels, fricatives, affricatives) .

For example, the word "9" ("DEVET") in over 60% cases is recognized as "5" ("PET"). The similar case is for word "2" ("DVA") which mostly refers to "0" ("NULA"). If we eliminate these critical words from the dictionary, recognition rate for unknown speakers is 88%, and for known speakers is 92% and that corresponds to recognition rate of existing isolated word speaker independent recognition systems.

4. CONCLUSIONS

A speech recognition interface for telephone IVR systems was developed for speaker independent isolated word recognition. Initial result show that the recognition accuracy is above 80% for independent speaker and 100% for dependent speaker recognition with carefully dictionary design. Experiments has shown that different words in the same dictionary which are phonetically similar have to be represented with greater number of reference templates.

To increase the recognition rate it has to avoid using short words (like "DVA" and "TRI" - longer utterances are more disriminable), words with similar sub-parts which are easily confused ("TRI" and "ČETIRI").

To reduce the dictionary size and the processing time, many linked dictionaries may be used in hierarchical structure to simulate larger vocabulary. This method is ideal for the menu driven interface applications which are wide used in IVR systems.

5. REFERENCE

[1] I. Kraljevski, D. Mihajlov, D. Gorgjevik, "Hybrid HMM/ANN System for Speech Recognition on Macedonian Language", ETAI 2000, V - National Conference, Ohrid, R. Macedonia. 2000. II-1: 1-7.

[2] S. N. Wrigley. Speech Recognition by Dynamic Time Warping
http://www.dcs.shef.ac.uk/~stu/com326/dtw.htm
[01/15/2002]

[3] Rabiner L.R., Levinson S.E., Rosenberg A.E., Wilpo J.G., "Speaker Independent Recognition of Isolated Word using Clustering Techniques", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. ASSP-27, 336-349.

[4] E. Dermatas, G. Kokkinakis, "Multiple Templates Algorithm for Speaker Independent Isolated Word Recognition Systems"