TRIPHONE BASED CONTINUOUS SPEECH RECOGNITION SYSTEM FOR TURKISH LANGUAGE USING HIDDEN MARKOV MODEL

Fatma Patlar
Istanbul Kültür University
Department of Computer Engineering
34156 Bakirkoy / Istanbul
fatmapatlar@yahoo.com

Akhan Akbulut
Istanbul Kültür University
Department of Computer Engineering
34156 Bakirkoy / Istanbul
a.akbulut@iku.edu.tr

ABSTRACT
This paper introduces a system which is designed to perform a relatively accurate transcription of speech and in particular, continuous speech recognition based on triphone model for Turkish language. Turkish is generally different from Indo-European languages (English, Spanish, French, German etc.) by its agglutinative and suffixing morphology. Therefore vocabulary growth rate is very high and as a consequence, constructing a continuous speech recognition system for Turkish based on whole words is not feasible. By considering this fact in this paper, acoustic models which are based on triphones, are modelled as five state Hidden Markov Models (HMM). Mel-Frequency Cepstral Coefficients (MFCC) approach was preferred as the feature vector extraction method and training is done using embedding training that uses Baum-Welch re-estimation. Recognition is implemented on a search network which can be ultimately seen as HMM states connected by transitions and Viterbi Token Passing algorithm runs on this network to find the mostly likely state sequence according to the utterance. Also to make a more accurate recognition bigram language model is constructed.

KEY WORDS
Continuous Speech Recognition, Triphone, Hidden Markov Model, Language Modelling, Bigram language model, Turkish

1. Introduction

Speech communication within human beings has played an important role because of its inherent superiority over other modes of human communication. As time passed, the need for better control of complex machines appeared and speech response systems have begun to play a major role in human-machine communication.

The use of voice processing systems for voices input and output provides a degree of freedom for mobility, alternative modalities for command and data communication and the possibility of substantial reduction in training time to learn to interface with complex system. All those characteristics of speech yield lots of positive advantages over other methods of human-machine interaction, when incorporated into an effective voice control or voice data entry system [1].

Most successful speech recognition systems today use the statistical approach. The Hidden Markov Models are used to model speech variations in a statistical and stochastic manner [2]. The idea is collecting statistics about the variations in the signal over a database of samples, and uses these to make a representation of the stochastic process. The speech signal carries acoustic information about the symbolic translation of its meaning. The acoustic information is embodied in HMMs and sound events are modelled by them. Besides the acoustic information, the sequence of the symbols also constitutes a stochastic process, which can also be modelled. Such statistical approaches to model the symbolic events are called language models.

There are few studies in Turkish speech recognition with respect to the most of the other languages. Especially the large vocabulary recognition field for Turkish is quite empty. The main reason for this is the difficulties caused by the Turkish language. Turkish is an agglutinative language which makes it highly productive in terms of word generation. This is an important problem for traditional speech recognizer which has a fixed vocabulary.

This paper represents an overview of our approach [3]. Section two explains the studies conducted so far. In section three, the details of our proposed Speech Recognition System is listed. Test results are given in section four and section five concludes the whole paper and gives some future notifications.

2. Background

It has been over 60 years since first works on the speech recognition and so much has been accomplished not only in recognition accuracy but also in performance factor.

For example in isolated word recognition, vocabulary sizes have increased from ten to several hundred words. Highly confusable words have been distinguished, improved adaptation to the individual speaker dependent and speaker independent systems have been developed, the telephone and noisy, distorting channels have been effectively used and effect of other environmental...
conditions like vibration, g-forces, and emotional stress have also been explored [4].

As a result, there are now several research prototype spoken language systems that support limited interactions in domains such as travel planning, urban exploration, and office management. These systems operate in near real life, accepting spontaneous, continuous speech from speakers with no prior enrolment; they have vocabularies of 1000-2000 words, and an overall correct understanding rate of almost 90%.

If we look at the speech recognition from Turkish Language perspective, there are only a few studies on Turkish speech recognition. Most of the studies were being done on isolated word recognition, where only a word is recognized at each search step [5] [6] [7] [8]. Similar or same methods used in English speech recognition generally also can be used in Turkish recognition. The acoustic modeling is almost universal for all languages, since the methods that are used yield similar results [9] [10]. However, the classical language modeling techniques cannot be applied successfully to Turkish.

Turkish is an agglutinative language. Thus, great number of different words can be built from a base by using derivations and inflections. The productive derivational and inflectional structure brings the problem of high growth rate of vocabulary. Furthermore, Turkish is a free word order language. So the language models that are constructed using the co occurrences of words do not perform well.

3. Approach

In this section, the experimental work on continuous speech recognition system for Turkish Language based on Triphone Model will be presented.

MATLAB is used for data Data Preparation and Pre-Processing steps and Hidden Markov Toolkit (HTK) is used for Feature Extraction, Training and Recognition steps.

Block Diagram of our speech to text system can be given as shown in Figure 1.

3.1 Data Preparation

Two different databases are used, one of them is more commonly formed TURTEL speech database which are collected at the acoustics laboratory of TUBITAK-UEKAE (National Research Institute of Electronics & Cryptology, The Scientific & Technical Research Council of Turkey), that is used for speaker independent system tests and the other one is weather forecast reports database that is used for speaker dependent system tests.

3.2 Data Preparation

Preprocessing involves three steps: noise cancelling, preemphasis and voice activation detection. Noise-cancelling reduces unwanted ambient sounds by a low pass filter which is a filter that passes low-frequency signals. As the part of preemphasis, the first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies [11].

This is usually done by a highpass filter. The most commonly used filter type for this step is the FIR filter. The filter characteristic is given by the equation below.

\[ H(z)=1-a z^{-1} \]  

The value (a) is chosen to be approximately 0.95. There are two reasons behind this. The first is that to introduce a zero near \( z=1 \), so that the spectral contributions of the larynx and the lips have been effectively eliminated. The second reason, if the speech signal is dominated by low frequencies, preemphasis is used to prevent numerical instability [12].

For our system, we choose \( a=0.95 \). We found an optimum value by trial and error.

An important problem in speech processing is to detect the presence of speech in a background noise (Voice Activation Detection). This problem is often referred to as the end point location problem [13] [14].

The accurate detection of a word start and end points means that subsequent processing of the data can be kept to a minimum. The energies of each block is calculated first using sum of square energy algorithm.

\[ E = \sum_{i=1}^{N} x(i)^2 \]  

Assuming that there is no speech in the first few frames, of recording, the average of the first few frames, give the maximum energy and the minimum energy which are used calculates for cutting down unvoiced parts of speech.
3.3 Feature Extraction

The purpose of feature extraction is to convert the speech waveform to some type of parametric representation (at a considerably lower information rate). The speech signal is a slowly time-varying signal (it is called quasi-stationary) [15]. When examined over a sufficiently short period of time (typically between 20 and 100 ms), its characteristics are fairly stationary.

A wide range of possibilities exist for parametrically representing the speech signal for the recognition task, such as Linear Prediction Coding (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and others. MFCC’s are based on the known variation of the human ear’s critical bandwidths with frequency. MFCC’s apply Mel scale to power spectrum of speech in order to imitate human hearing mechanism. MFCC is the best known and most popular, and this feature has been used in this work.

3.3.1 Frame Blocking

Speech signals change in every few millisecond, because of this speech should be analyzed in small duration frames.

In this step the reemphasized speech signal is blocked into frames of N=200 samples (25ms) and each frame overlaps with the previous frame by a predefined size that in our case separated by M=80 samples (10 ms, typical to be stationary signal).

3.3.2 Windowing

Choosing a window type is an important factor to process signal correctly. Two competing factors exist in this choice. One of them is smoothing the discontinuity at the window boundaries and other factor is not to disturb the selected points of the waveform. Typical window types are Rectangular, Hamming and Hanning.

A more common window used in MFCC extraction is the Hamming window, which shrinks the values of the signal toward zero at the window boundaries, avoiding discontinuities.

3.3.3 Fast Fourier Transform

The next step is to extract spectral information for our windowed signal; we need to know how much energy the signal contains at different frequency bands.

In here Fast Fourier Transform (FFT) is used to convert speech frame to its frequency domain representation, the short term power spectrum is found.

In order to lead FFT do it job efficiently, N, point number, has to be a power of 2. Here N is selected as 256 points (200 samples ≈ 256). Fast Fourier Transform is used with zero padding (256=200 samples+56 zero). Zero-padding append an array of zeros to the end of the input signal before FFT.

3.3.4 Mel Filter Bank

Perceptual filter bank (a Mel-scale filter bank Eq.(3)) is used to approximate the human ear’s response to speech (in an attempt to receive only relevant information). Due to the overlapping filters, data in each band highly correlated. Filters are used to emphasize some of the frequency contents in power spectrum of the speech like ear does. More filter in the bank process the spectrum below 1 kHz since the speech signal contains most of its useful information such as first formant in lower frequencies.

$$f_{mel} = 2595 \log_{10}(1 + \frac{f}{700})$$

The spacing between the filters in mel scale is computed using:

$$\Delta \theta = (\theta_{max} - \theta_{min})/(M + 1)$$

Center frequencies of the filters are found using:

$$\theta_c(m) = m, \Delta \theta, \quad m = 1, 2, \ldots, M$$

After converting these center frequencies to Hertz, the filter are formed using the formulae below:

$$h(k,m) = \begin{cases} 
0 & \text{for } f(k) < f_c(m-1) \\
\frac{f(k) - f_c(m)}{f_c(m+1) - f_c(m-1)} & \text{for } f_c(m-1) \leq f(k) < f_c(m) \\
\frac{f(k) - f_c(m)}{f_c(m+1) - f_c(m)} & \text{for } f_c(m) \leq f(k) < f_c(m+1) \\
0 & \text{for } f(k) \geq f_c(m+1) 
\end{cases}$$

In this work 26 triangular filters are formed to gather energies from frequency bands. They are all linearly spaced along Mel scale. According to the cut of 120hz-3400hz, 14 of these filter are below 1000hz and rest of them are above 1000hz.

3.3.5 Discrete Cosine Transformation

After filtering, discrete cosine transform (DCT) is applied to the resulting log filter-bank coefficients to compress the spectral information into lower order ones, and also to de-correlate them. Generally the first 12 coefficients
which we will get from DCT are enough for MFCC purposes. Higher cepstral coefficient can be used to determine pitch information.

After this process we have Mel spectrum of the signal, next step is determining the Mel cepstral coefficients. To do this, logarithms of filterbank amplitudes are taken, and then using DCT, 12 DCT coefficients and 1 energy coefficient are calculated (39 size vector with 12 delta coefficients plus 1 energy and 13 double deltas). These coefficients are calculated in Mel scale and they are FFT based cepstral coefficients.

The main reason of preferring DCT rather than IFFT is IFFT requires complex arithmetic calculations compared to DCT.

3.4 Training and Recognition

There are advanced techniques in speech recognition such as Dynamic Time Warping (DTW), the Hidden Markov Modelling (HMM) and Artificial Neural Network (ANN) techniques.

Acoustic models are formed as Hidden Markov Models. The parameter estimation is implemented using a computationally efficient algorithm known as Baum Welch re-estimation (also referred to as the Forward Backward algorithm). Baum-Welch training can be viewed as providing the system a capability to make soft decisions.

Training process start with initialization of monophone models; global mean and variance of the training data is assigned to each state by using flat start method whereby all models are given the same initial parameters. After initialization, triphone models are initialized using monophone models and retrained using embedded training method. In embedded training composite HMM is created by concatenating sub word models that corresponds to the sub words in the record text files.

To limit the complexity of the triphone models and avoid the sparse data problem in acoustic training, decision tree based state clustering is used. Also decision tree clustering is used to increase robustness.

After preparation of training modules, embedded training is performed.

4. Test Results

In recognition experiments, word accuracy of speaker independent system has been measured as 59-63 percent. After finding optimum value for decision tree pruning factor by try outs, system tests have been repeated again by using the language model and the optimum pruning factor. These adjustments improved the performance by 30-33 percent and word accuracy has reached to 92-93 percent for all tests.

While the word accuracy of the speaker dependent system tests on the single person database is between 89-93 percent, usage of the language model and the optimum decision tree pruning factor has resulted with an increase in the performance and the word accuracy has reached to 95-97 percent.

5. Conclusion

In this work, a triphone model based, continuous speech recognition system is developed for Turkish language. Besides acoustic model which is the core part of the system, a statistical language model is also used to improve recognition performance.

Turkish has an agglutinative morphology with productive inflectional and derivational suffixations. Because of these productive suffixations the number of words in vocabulary is very high. This means that a whole word recognition approach cannot be feasible for Turkish language. Therefore, triphones are used as the smallest unit in the acoustic model and they are modelled by 5-state left-to-right Hidden Markov Models which are also called Bakis model.

When designing large vocabulary cross-word triphone system it is unavoidable that there will be triphones which has no examples in the training data. To overcome this problem a decision tree approach is also investigated and implemented.

Training process is done with embedded training using Baum-Welch algorithm. By this way parameter estimation process for all models occurred in parallel.

A bigram statistical language model is also used in the recognition and proposed system tested on two different speech databases, TURTEL Speech Database and Weather Forecast Reports Database, with different pruning factors and also with/without using a language model.

The tests with these databases have shown that for large vocabulary cross-word recognition systems, language models can provide a good amount of increase in the recognition accuracy.

The experimental results had shown that the proposed system worked well on both speech databases and had accuracy between 59-96% in correctly identifying a speech sentence.

6. Future Work

We considered that the language model can be a good topic to make improvements. Using a trigram language model in the place of a bigram language definitely provides an increase in the recognition performance.

Another topic for future work can be robustness against noise and distortion. Some of the noise compensation methods for Hidden Markov Models in [16] can be investigated and put into practice for this work.
Acknowledgements

This paper was prepared in collaboration with Asst.Prof. Dr. Ertugrul Saatci. We would like to thank The Scientific and Technical Research Council of Turkey (TUBITAK, UEKAE) for supporting us.

References