VERİ BİLİMİ DERGİSİ
www.dergipark.gov.tr/veri

# Turkish Speech recognition using Mel-frequency cepstral coefficients(MFCC) and Hidden Markov Model (HMM)

Mustafa Cumaah AHMED[1], Hasan Erdinç KOÇER[2]*

[1]*Konya Technical University, Graduate Education Institute, Computer Engieneering, Konya*
[2]*Selçuk University, Technology Faculty, Electrical & Electronics Engineering, Konya*

**Makale Bilgisi**

*Başvuru:*
*15/11/2019*
*Kabul:*
*23/12/2019*

## Abstract

In this paper, a new Turkish spoken number recognition system proposed. The Mel-frequency cepstral coefficients (MFCC) algorithm used as a feature extraction method, the Gaussian Hidden Markov model, used for numbers phonemes modeling where each number has a Markov model. The system trained on a dataset collected from 20 subjects that includes 7 females and 13 males. Each one says the Turkish numbers from “zero” to “ten”. Audio files sampled at 8000Hz at each second and each file has one-second length and recorded in an isolated environment. We tested the system using random records for different people. The training files include 220 audio record and testing files include 18 audio record. The system achieves %83.3 accuracy, %86 precision, and %83 recall rates.

***Keywords:*** *Hidden Markov Model, Mel-Frequency Cepstral Coefficients, Turkish Speech Recognition*

## 1 Introduction

Speech recognition is the process that machines recognize the words spoken by the human and one of the most investigated human-computer interfaces over the last decades [1][2]. Automatic Speech recognition (ASR) system works by identifying the uttered word according to prior knowledge of the features extracted from the spoken words speech signal using a feature extraction method. Theoretically, the speech signal contains a considerable variability caused by the background noise, emotions, and expression. However, Automatic speech recognition uses the feature extraction to reduce that variability by eliminating the effects of pitch, fundamental frequency, and the amplitude of the excitation signal. The main reason that ASR systems compute the short-term spectrum is that the human ear cochlea performs a quasi-frequency analysis which is a non-linear frequency scale analysis approximately up to 1000Hz. So it is necessary to perform frequency warping. Recently, the researchers proposed automatic speech recognition approaches based on these feature extraction methods; LPC (Linear Predictive Coding) which is a most potent method for encoding speech signal at a low bit rate, the concept of LPC algorithm is that the specific speech sample at a specific time can be estimated as a linear combination of the past samples. LPC algorithm has many types such as; Voice-excitation LPC, Residual Excitation LPC, Pitch Excitation LPC, Multiple Excitation LPC, Regular Pulse Excited LPC and Coded Excited LPC. One of the most used methods for feature extraction is the Mel Frequency Cepstral Coefficients (MFCC) which can be considered as a standard method for speech signal analysis in the automatic speech recognition systems [3]. Most of the proposed MFCC-based systems used about 20 coefficients and often 10-12

***** Corresponding author: ekocer@selcuk.edu.tr coefficients for effective speech signal coding [4]. The MFCC algorithm is sensitive to the noise because it depends on the spectral form, so to avoid this problem most of the methods employ information in the periodicity of speech signals and also contains aperiodic signals [5]. The idea behind the MFCC algorithm is that the non-linear frequency scale used an estimate to Mel-Frequency scale which is a linear for frequencies below 1 kHz and logarithmic for above, this concept inspired by the biology because of the human auditory system becomes less selective for the frequencies above 1 kHz. The features extracted using MFCC correspond to the cepstrum of the log filter bank energies. On the other hand, the power spectral analysis techniques such as Fast Fourier transform (FFT) sometimes applied in many ASR systems where the frequency content of the signal over time is described by the power spectrum. In this research paper, we introduce a speech recognition system based on MFCC features and Hidden Markov model for Turkish numbers (0-10). We collect a speech dataset from 20 subjects each one utter the numbers from zero to ten in the Turkish language in an isolated recording environment. The collected dataset containing 220 files sampled at 8 kHz and one second long with 13 male and seven female subjects. The hidden Markov model used to model each phoneme for the numbers features. The recorded files contain more than one linguistic collected from foreigners.

## 2 Methods

### 2.1 Mel frequency cepstral coefficients(MFCC)

Mel-frequency cepstrum is a short-term representation of the power spectrum of a sound signal. The MFC based on linear cosine transform of a log power spectrum on a nonlinear mel-scale of frequency. MFCC is the coefficients collected using MFC analysis [6] which are derived from a cepstral type representation of a speech signal. The difference between MFC and cepstrum is; the frequency bands spaced on mel-scale in MFC which inspired from the human auditory system's response closely more than linear frequency bands in the normal cepstrum. The process of MFCC features extraction shown in figure 1.

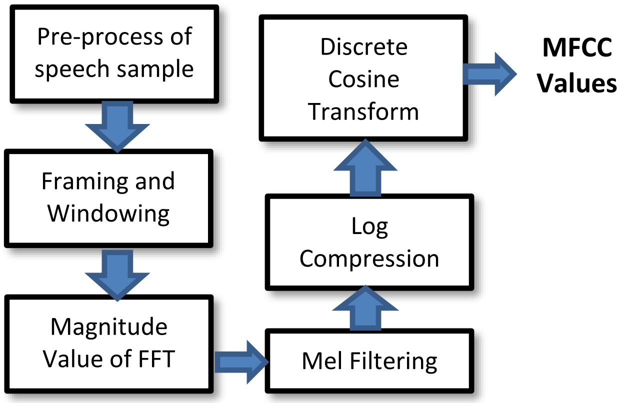


Fig. 1. MFCC algorithm Flow chart

The MFCC algorithm consists of seven steps to extract the features from the speech signal as follows:

1. Pre-process: The pre-process is a filter applied to the speech signal to amplify the higher frequencies. It is useful to balance the frequency spectrum because of the higher frequencies have small magnitudes compared to the lower frequencies. It is also used to avoid the numerical problems when the fast Fourier transform operation is applied and improve the signal-to-noise ratio (SNR). The pre-emphasis applied to x signal as follows :

$$y(t) = x(t) - \alpha\, x(t-1)$$

The typical values of the filter coefficient α are between 0.95 to 0.97.

The pre-emphasis filter has a significant effect on the modern systems because it can be achieved merely by using mean normalization to avoid the numerical issues of the FFT in the modern implementations.

2. Framing: After pre-emphasis, framing step used to split the signal into short-time frames. It used because of the frequencies in the signal changes over the time. The reason behind the framing is that it does not make sense to apply FFT on the entire speech signal to avoid frequency contours loss of the signal over the time. The typical frames period is between 20ms-40ms with 50% overlap between sequential frames. By using framing, we can suppose that the frequencies in the speech signal are fixed over a short period so we can obtain a good

estimation of the frequency contours by doing Fourier transform.

3. Windowing: After framing the speech signal hamming window applied on each frame as follow:

$$w_n = 0.54 - 0.46cos\left(\frac{2\pi \boldsymbol{n}}{\mathrm{N}-1}\right)$$

Where, 0≤n≤N−1, N is the window length as shown in figure 2.

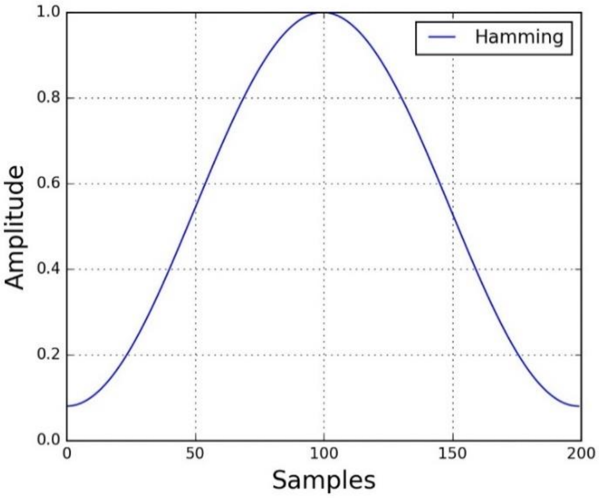


Fig 2. Hamming Window

The reason for applying the hamming window is to prevent the assumption made by the FFT that the signal is constant and to reduce spectral leakage.

4. FFT: An N-point FFT performed on each frame to calculate the frequency spectrum. Typically N is 256 or 512. After extracting FFT from each frame, the power spectrum will be computed as follows:

$$P = \frac{|\,FFT(x_i)|^2}{N}$$

Where $x_i$ is the $i^{th}$ frame of signal x.

5. Filter banks: The filter bank is the final step and computed by applying triangular filters on mel-scale typically 40 filters. The mel-scale filter bank mimics the perception of the human ear for sounds. The converting between hertz and mel can be done as follows:

$$m = 2595log_{10}\left(1 + \frac{f}{700}\right)$$

$$f = 700\left(10^{m/2595} - 1\right)$$

Each filter is triangular having response one at the center frequency and decreases to zero linearly.

6. MFCC Features: The filter banks output is highly correlated which is a problem for many machine learning algorithms, for this reason, Discrete Cosine Transform(DCT) applied to decorrelate the filter bank coefficients and compress the filters representation. Typically in the automatic speech recognition system, 2-13 coefficients are used. Figure 3 shows MFCC features for spoken number "six" in Turkish and figure 4 shows features of spoken number "seven" in the Turkish language.

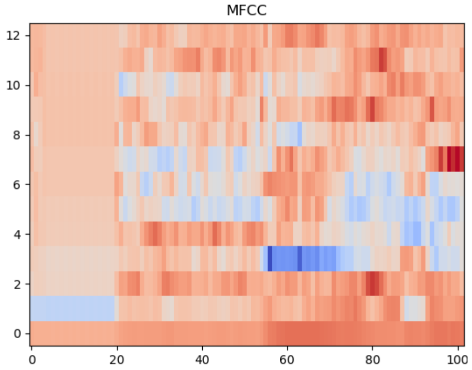


Fig 3. MFCC features of spoken number "six" in Turkish

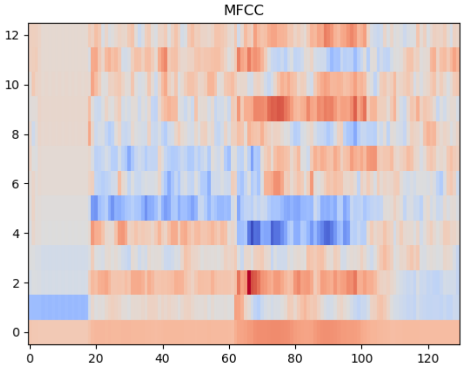


Fig 4. MFCC features of spoken number "seven" in Turkish.

### *Hidden Markov Model*

Hidden Markov model is a useful modeling method for time-varying spectral features which is the most used technique in modern speech recognition systems. It is virtually used almost in all ASR systems as a basic framework over the last decades[7,8,9]. The modern HMM-based speech recognition systems founded by the group of researchers at Carnegie Mellon University and IBM in the 1970s. The proposed use of the discrete density HMMs[10,11,12]. The continuous density HMMs for speech recognition introduced by Bell labs [13,14,15].

The principal of HMM-based speech recognizer shown in figure 5. The input is the speech signal of the spoken word which converted to a feature sequence using MFCC algorithm.

$$Y_{1:T} = \{y_1 \ldots\ldots y_T\}$$

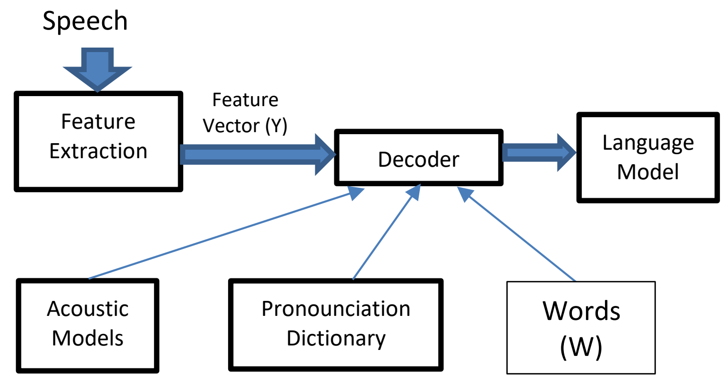


Fig 5. HMM-based Speech recognizer

The decoder tries to a sequence of words.

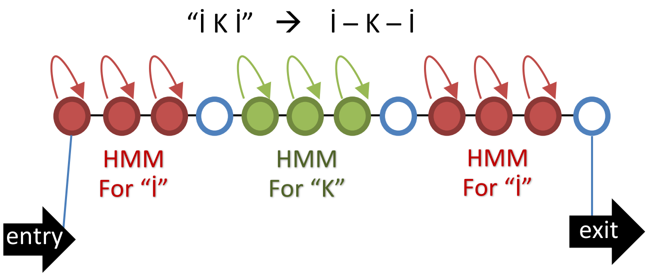
$$w_{1:L} = \{w_1 \ldots\ldots w_L\}$$

Which is most likely to have generated Y, as follows:

$$\widehat{w} = \underbrace{argmax}_{w} \{p(Y|w)P(w)\}$$

The likelihood p(Y |w) obtained by the acoustic model and P (w) determined by the language model. Each unit of speech represented as a phone by the acoustic model for example cat represented as /c/, /a/, /t/.

For any given word w the corresponding model synthesized by concatenating the phonemes to make the words as the dictionary. The parameters of the models estimated from the training data that contains the uttered words speech signal. The language model usually is the N-gram model which each word probability conditioned only on its N – 1 predecessors.



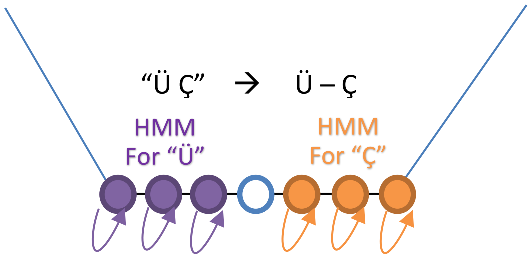


Fig 6. Example of Hidden Markov modelling for number "iki" and "üç"

### *2.3 DATASET*

We collected the dataset from 20 subjects; 13 male and seven female subjects, all records are sampled at 8000kHz at second and one-second length. The dataset collected from Turkish subjects and foreigners, the linguistic differs in some records. For the testing data, random records collected from female and male subjects vary from the training data.

Table 1. Dataset Information

| Gender | Male | Female |
|---|---|---|
| *Sampling rate* | *8000kHz* | *8000kHz* |
| *Files per subject* | *11* | *11* |
| *Subjects* | *13* | *7* |
| *Record length* | *1s* | *1s* |
| *Channel* | *Mono* | *Mono* |

### *2.4 PROPOSED SYSTEM*

We proposed an automatic speech recognition for Turkish spoken numbers from "zero" to "ten". The system is based on Mel-frequency cepstral coefficients (MFCC) as a features extraction. We used 13 coefficients for our system. Hidden Markov Model (HMM) with Gaussian emissions used for modeling [16]. Figure 7 shows the block diagram of the ASR system.

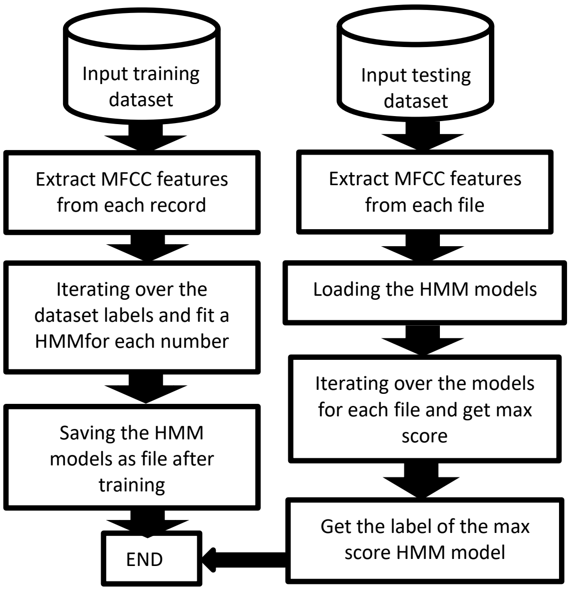

Input training dataset
Extract MFCC features from each record
Iterating over the dataset labels and fit a HMMfor each number
Saving the HMM models as file after training
END
Input testing dataset
Extract MFCC features from each file
Loading the HMM models
Iterating over the models for each file and get max score
Get the label of the max score HMM model

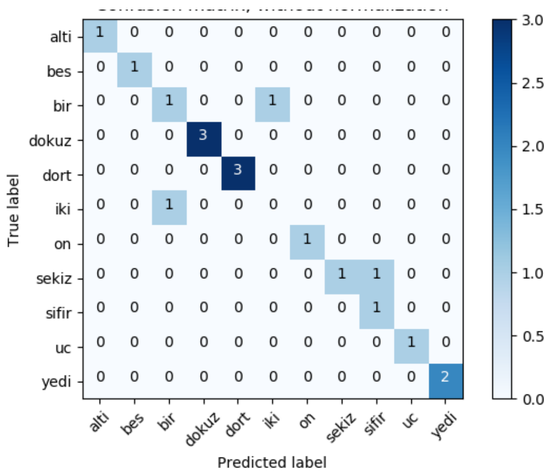
Fig 7. The block diagram of the proposed system.

## 3 Results

The implementation of the program is done using Python programming language and HMMlearn Hidden Markov model library. The training is done using the dataset that containing 20 subjects and 220 total files, 13 MFCC coefficients extracted from each file. Each number features modeled using Hidden Markov model with Gaussian emission. The HMM model has 11 states and 1000 iteration. We test the program using a test set containing 18 records mixed with female and male subjects, it achieved %83.3 recognition accuracy, %86 precision, and %83 recall. Bellow table II shows the result. Figure 9 shows the confusion matrix.

Table 2. Results of MFCC-HMM ASR system

| Number | Precision | Recall | F1-score | Number of files |
|---|---|---|---|---|
| Zero | %50 | %100 | %67 | 1 |
| One | %50 | %50 | %50 | 2 |
| Two | %0 | %0 | %0 | 1 |
| Three | %100 | %100 | %100 | 1 |
| Four | %100 | %100 | %100 | 3 |
| Five | %100 | %100 | %100 | 1 |
| Six | %100 | %100 | %100 | 1 |
| Seven | %100 | %100 | %100 | 2 |
| Eight | %100 | %50 | %67 | 2 |
| Nine | %100 | %100 | %100 | 3 |
| Ten | %100 | %100 | %100 | 1 |
| Total | %86 | %83 | %83 | 18 |



| | alti | bes | bir | dokuz | dort | iki | on | sekiz | sifir | uc | yedi |
|---|---|---|---|---|---|---|---|---|---|---|---|
| alti | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bes | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bir | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| dokuz | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| dort | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| iki | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| on | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| sekiz | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| sifir | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| uc | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| yedi | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |

Fig 9. The Recognition confusion matrix.

## 4 Conclusion

The automatic speech recognition system is proposed in this research paper based on mel frequency cepstral coefficients(MFCC) and Hidden Markov model with Gaussian emission. The ASR system shows a good result on the collected dataset of the spoken Turkish numbers. Each number have been modeled at the training time and the trained models saved as a file. The program loads the trained HMM models and iterate over the testing set and calculate the maximum score of all of HMM models and get the labels. Accuracy, precision, and recall metrics are calculated. This research paper shows a potential to use the MFCC and Hidden Markov model with Gaussian emissions to build a large scale vocabulary Turkish speech recognition system in the future.

## References

[1] Rabiner, Lawrence R., and Biing-Hwang Juang. Fundamentals of speech recognition. Vol. 14. Englewood Cliffs: PTR Prentice Hall, 1993.

[2] Deller, John R., John HL Hansen, and John G. Proakis. "Discrete-time processing of speech signals." (2000): 595-602.

[3] Motlıcek, Petr. Feature extraction in speech coding and recognition. Technical Report of PhD research internship in ASP Group, OGI-OHSU, http://www. fit. vutbr. cz/~ motlicek/publi/2002/rep ogi. pdf, 2002.

[4] Hagen, Andreas, Daniel A. Connors, and Bryan L. Pellom. "The analysis and design of architecture systems for speech recognition on modern handheld-computing devices." Proceedings of the 1st IEEE/ACM/IFIP international conference on Hardware/ Software codesign and system synthesis. ACM, 2003.

[5] Ishizuka, Kentaro, and Tomohiro Nakatani. "A feature extraction method using subband based periodicity and aperiodicity decomposition with noise

robust frontend processing for automatic speech recognition." Speech communication 48.11 (2006): 1447-1457.

[6] Xu, Min, et al. "HMM-based audio keyword generation." Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2004.

[7] G. Evermann, H. Y. Chan, M. J. F. Gales, T. Hain, X. Liu, D. Mrva, L. Wang,and P. Woodland, "Development of the 2003 CU-HTK conversational telephone speech transcription system," in Proceedings of ICASSP, Montreal, Canada, 2004.

[8] S. Matsoukas, J.-L. Gauvain, A. Adda, T. Colthurst, C. I. Kao, O. Kimball, L. Lamel, F. Lefevre, J. Z. Ma, J. Makhoul, L. Nguyen, R. Prasad, R. Schwartz, H. Schwenk, and B. Xiang, "Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1541–1556, September 2006.

[9] H. Soltau, B. Kingsbury, L. Mangu, D. Povey, G. Saon, and G. Zweig, "The IBM 2004 conversational telephony system for rich transcription," in Proceedings of ICASSP, Philadelphia, PA, 2005.

[10] J. K. Baker, "The Dragon system — An overview," IEEE Transactions on Acoustics Speech and Signal Processing, vol. 23, no. 1, pp. 24–29, 1975.

[11] F. Jelinek, "Continuous speech recognition by statistical methods," Proceedings of IEEE, vol. 64, no. 4, pp. 532–556, 1976.

[12] B. T. Lowerre, The Harpy Speech Recognition System. PhD thesis, Carnegie Mellon, 1976.

[13] B.-H. Juang, "On the hidden Markov model and dynamic time warping for speech recognition — A unified view," AT and T Technical Journal, vol. 63,no. 7, pp. 1213–1243, 1984.

[14] B.-H. Juang, "Maximum-likelihood estimation for mixture multivariate stochastic observations of Markov chains," AT and T Technical Journal,vol. 64, no. 6, pp. 1235–1249, 1985.

[15] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition," Bell Systems Technical Journal, vol. 62, no. 4,pp. 1035–1074, 1983.