Uttered Kurdish digit recognition system

Saman Muhammad Omer Koya University- Faculty of Engineering- Department of Software Engineering.
Email: saman.muhammad@koyauniversity.org.

Jihad Anwar Qadir University of Raparin - College of Basic Education - Department of Computer Science.

Email: jihad.qadir@uor.edu.krd.

Zrar Kh. Abdul [1] Charmo University - College of Medicals and Applied Sciences-Department of Applying Computer.
[2] University of Halabja- College of science- Department of Computer science.
Email: zrar.abdul@charmouniversity.org

Abstract:

Speech recognition is a crucial subject in human computer interaction area. The ability of a machine to recognize words and phrases in spoken language is speech recognition and then convert them to a machine-readable format. Digit recognition is a part of the speech recognition system. In this paper, three spectral based features including Mel Frequency Cepstral Coefficient (MFCC), Linear predictive coding (LPC) and formant frequencies are proposed to classify ten Kurdish uttered digits (0-9). The features are extracted from entire speech signal, and feed a pairwise SVM classifier. Experiments including each individual feature and different forms of fusion are conducted and the results are shown. The fusion of the features significantly improves the result and shows that the different features carry complementary information. The proposed model is experimented on the dataset that have been collected in Kurdistan.

Key words: Speech recognition, MFCC, LPC, Formant frequencies, uttered digits, SVM

1. Introduction:

Speech is an important way of communication among human, and it is the most natural and efficient form of exchanging information among them. There are still ongoing investigations among researchers to improve the interconnection between human and computer. In (Yu and Deng, 2016, p.29) proposed a system to be able to compile human speech to computer info in order to understand and recognize speech signal. Therefore, speech recognition can be defined as the process of converting speech signal to a sequence of words. However, there is still a performance gap between Human Speech Recognition (HSR) ability and Automatic Speech Recognition (ASR) performance.

There are two well-known approaches in speech processing, utterance (or words) isolating and spontaneous speech recognition (Gaikwad et al., 2010, p.16). System based on utterance isolated speech is able to recognize a word or utterance at a time. while systems based on spontaneous speech is able to process a continuous speech and the system allow user to speak more freely and naturally.

We can find in the previous works that, some researchers are focusing on the speech recognition based on the features extracted from an isolated word. In (Sakoe et al., 1989, p.439) proposed a new speech recognition model using time-sequence-structure for a neural network known as Dynamic Programming Neural Network (DNN). Speaker-independent isolated Japanese digit words were recognized, as the primary experimentation. The highest recognition accuracy was achieved as 99.3%. (Bilginer Gülmezoğlu, 1999, p.620) developed two theories by considering two optimization criteria, DTW and NN methods, applied to both the training set and the test set and also LPC, CEP, ECEP, and R- MCEP are used as feature extraction method. The low recognition rate was achieved with NN approach but when R-MCEP parameters are used the highest recognition rate was achieved. In (Wijoyo and Wijoyo, 2011, p.29) developed a speech recognition system on a mobile robot for controlling movement of the robot by using LPC and ANN method on this system, the highest average recognition rate that was achieved by the system was 91.4%.

The other side, spontaneous speech recognition studies are interesting in the last decades. In (Lee and Hon, 1989, p.1641) Proposed speaker-independent phone recognition system by extending a common Hidden Markov Modelling (HMM). Using multiple codebooks of different Linear Predictive Coding (LPC) parameters and discrete HMM's, they reported their results as 73.80% accuracy for the speaker-independent phone recognition. They also introduced the co-occurrence smoothing technique which enables accurate recognition even with very limited training data. (Thiang, 2007, p.1193) studied a speech recognition system on MCS51 microcontroller that recognize the word used as the command for controlling movement of a wheelchair. in the proposed system, two approaches are implemented namely, Linear Predictive Coding (LPC) combined with Euclidean Squared Distance(ESD) approach and Hidden Markov Model (HMM) Segmentation and Centroid. The best achieved recognition rate using LPC-ESD method was 78.57%, and 32.86% was achieved using HMM-Segmentation and Centroid approach. (Thiang and Wijaya, 2009, p.347)Implemented a speech recognition system on ATmega162 microcontroller, which is applied for controlling movement of a mobile robot using LPC and HMM. The highest achieved recognition rate was 87%. MFCC, LPC and formant features are one of the most common features that have been used in the speech processing field.

DOI:https://doi.org/10.26750/Vol(6).no(2).paper 5 http://journal.uor.edu.krd/index.php/JUR Vol.6. No.2, September.2019

(Al-Talabani et al., 2017, p.20) developed a system for dialect and language recognition using these mentioned features in addition one dimensional Local Binary Pattern (LBP). Recently, Convolutional Neural Network (CNN) has been fed by the MFCC frames for Kurdish speaker identification, and the result shown that the MFCC frames can improve the performance of the speaker identification (Abdul, 2019, p.566).

Through the literature review, we could not find any published work to develop any Kurdish dialect speech recognition system. In this paper, a system has been proposed to recognize ten uttered digits in Kurdish language from zero to nine. The suggested features extracted from the recorded speech which are LPC, MFCC, and Formant. The various sets of features are fused in the feature level (concatenated). In the classification mode, pairwise SVM has been used, with all of the suggested sets of features.

The rest of the papers include a section about the background of the feature extraction and classification level, followed by a data acquisition description in the section three. The fourth section is about the methodology of the conducted experiments, and finally the result and discussion and the conclusion come is presented in the fifth and the sixth sections respectively.

2. Background

This section focuses on the background of the features and classification technique used in this work. The presented well-known Features in the ASR area which are used in this study are: Mel Frequency Cypstrem Coefficients (MFCC), Linear Predicting Coefficients (LPC) and formant frequencies. While SVM is the classification technique used which will be presented in an independent section.

2.1 MFCC Feature

One of the well-known and used features in speech analysis like voice recognition, speaker recognition and gender identification is the MFCC. In the recent years, it has been used in numerous applications such as the bio-medical area and for the diagnosis of the child's body through the voice while crying (Gupta et al., 2013, p.101). The MFCC can be calculated based on short-term analysis. Thus, then MFCC is calculated for each overlapped frame with length of 30ms with 15ms overlap. MFCC captures the important characteristic of phonetic in speech and shows the shape of the vocal tract manifests itself in the envelope of the short time power spectrum (Muda et al., 2010, p.1003).
The computation of MFCC are shortened into some steps that mentioned below and shown in the figure (1):

1. Frame the signal into short frames.
2. Calculate the power spectrum.
3. Finding the mel filterbank to the power spectra
4. Take the logarithm of all filter bank.
5. Take the IDCT of the log filter bank.
6. Keep DCT coefficients 2-13, discard the rest.

DOI:https://doi.org/10.26750/Vol(6).no(2).paper 5 http://journal.uor.edu.krd/index.php/JUR Vol.6. No.2, September.2019

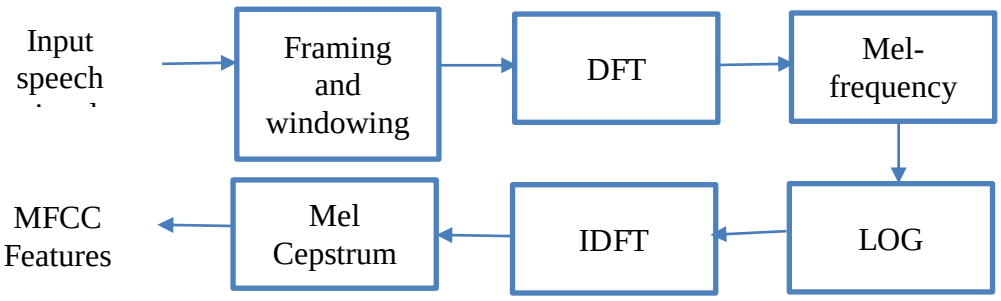

Figure 1 MFCC process

2.2 Formant Frequency

Formant frequencies are the spectral peaks of sound spectrum of the speech signal. The formant frequencies refer to the acoustic resonance of the human vocal tract and often measured as amplitude peaks in the frequency spectrum of the sound wave. It is well-known to separate the vowel phones, and has a good reputation for analysing emotional state of a person (Dave, 2013, p.1). In this paper, 12 peaks are extracted as Formant Frequency (4 formant frequencies, 4 magnitudes and 4 delta of magnitudes) then fed it to the Support Vector Machine (SVM).

2.3 LPC

Another powerful feature used in speaker recognition, is the Linear Predicting Coding (LPC), which can determine the parameters of speech signal and provides precise approximation of speech parameters. LPC is estimated by the following steps which are blocking, windowing, auto correlation and Linear prediction respectively (Furui, 1991, p.505) (Dave, 2013, p. 1)as shown in figure (2)

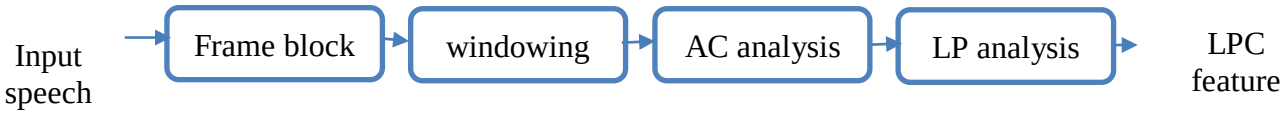

Figure 2 : LPC process

2.4 Support Vector machine

Classification is an important step in any machine learning system and used in various fields as they help us to make decisions about categorizing a data. One of the supervised methods is called Support Vector Machine (SVM) which is one of the most robust and successful classification method. The basic idea of SVM is to minimize the margin between two separated hyperplanes. However, SVM can handle binary classification problem. For multiclass problems different approaches need to be followed to perform the classification procedure. Pairwise SVM (followed in this study) which builds a machine for each pair of classes is one of the adopted approaches (Nath et al., 2014, p.554).

DOI:https://doi.org/10.26750/Vol(6).no(2).paper 5 http://journal.uor.edu.krd/index.php/JUR Vol.6. No.2, September.2019

Data acquisition

The dataset used in this study is designed and collected in Bosken primary School for this work on December 2018. All rules for collecting standard data are considered such that variance age and gender. The numbers of the subjects were 14 (7 males and 7 females). Their ages are seven (2 children), eight (4 children), nine (5 children), ten (2 children) and eleven (1 child) years old. The data collection procedure was to record the subjects' speech repeatedly five times saying words zero to nine. The total number of the recorded sounds was 700; 14 subjects *5 trail*10 classes (numbers)=700 sound clips. Eventually, the sounds were kept to be used in the (.wav) format in which each file contains one saved sound of an uttered number and then used as a dataset for this study.

Methodology

The conducted experiments in this study are designed by extracting three sets of features, including 12 MFCC, 12 LPCs and 12 formant frequencies based features (4 formant frequencies, 4 magnitude and 4 delta of magnitudes) which are all calculated and extracted from frame of the signal as the signals are divided into several frames with length 30 ms. Then all techniques (MFCC , LPC and Formant Frequency) are applied onto the frames to extract features. The mean and the standard deviation of the features over all of the frames are computed. 36 mean and 36 standard deviation features are computed (12 for each MFCC feature, 12 of the LPC features and 12 for each formant frequency).

In the classification level, we proposed pairwise SVM techniques with SMO optimization method. A total of 45 machines are trained (a combination of 10 class to each pair =10!/2!(10-2)!=45), then the final decision is done by majority voting. The experiments are conducted by feeding the models by each individual set of feature (MFCC, LPC, and Formants), then these features are concatenated into various sets. The results of all of the feature sets are presented as in the next section.

Result and Discussion

The experiments applied in this work are designed based on the extracted features (LPC, MFCC, and Formant frequencies). The result of these features feeding a pairwise SVM classifier with majority voting decision among the various machines is shown in table (1). The result of LPC shows significant improvement over MFCC and formant features with p value p=0.0001. However, the formant features have not shown significance improvement over the MFCC (p=0.06).

Table (1): The recognition rate for the proposed features

Feature extraction	**LPC**	**formant**	**MFCC**
Accuracy	87.86	86.14	85.86

DOI:https://doi.org/10.26750/Vol(6).no(2).paper 5 http://journal.uor.edu.krd/index.php/JUR Vol.6. No.2, September.2019

Some other experiments conducted to show how the various cases of fusion between the proposed features contribute in improving the accuracy rate. The result shown in table (2), clearly shows that the accuracy in all fusion case is significantly improving the accuracy rate compared to all of individual cases with p values <1*E-12. Moreover, among the fusion cases, there is significant improvement of LPC_formant and formant_MFCC against the LPC_MFCC with p value < 1*E-8. This might be an indication that MFCC and LPC are not adding extra separating information to each other since both, while formant features are looking at the problem from different points of view. Finally, a significant improvement over the two mentioned cases is performed with the LPC-MFCC-formant (P-Value<1*E-7).

Table (2): The recognition rate for the fused features

Feature extraction	**LPC_formant**	**LPC_MFCC**	**Formant_MFCC**	**LPC_formant_MFCC**
Accuracy	92.29	90.14	92.29	94.29

Conclusion

In this study we use three different acoustic features from speech signal to recognize uttered digits (0-9) in Kurdish language by children age (7 to 10). Although the basics of all of the proposed features (LPC, MFCC and formant) are cepstral, it shows that it carries some sorts of complementary information to each other. That is why the fusion among these features improves the results more than %6 of recognition rate. However, the adopted approach of having global features might be useful for limited dictionary, but not suitable enough as the dataset become larger.

Feature work

In future research we will attempt to use of local features, which can be extracted from the various frames instead of entire speech signal, could be useful even for limited number of words.

DOI:https://doi.org/10.26750/Vol(6).no(2).paper 5 http://journal.uor.edu.krd/index.php/JUR Vol.6. No.2, September.2019

Bibliography

Abdul, Z.K., 2019. Kurdish speaker identification based on one dimensional convolu- tional neural network 7, 566–572.

Al-Talabani, A., Abdul, Z., Ameen, A., 2017. Kurdish Dialects and Neighbor Languages Automatic Recognition. ARO-The Sci. J. Koya Univ. 5, 20–23.

Bilginer Gülmezoğlu, M., 1999. A Novel Approach to Isolated Word Recognition. IEEE Trans. Speech Audio Process. 7, 620–627.

Dave, N., 2013. Feature Extraction Methods LPC, PLP and MFCC 1, 1–5.

Furui, S., 1991. Speaker-dependent-feature extraction, recognition and processing techniques. Speech Commun. 10, 505–520.

Gaikwad, S.K., Gawali, B.W., Yannawar, P., 2010. A Review on Speech Recognition Technique. Int. J. Comput. Appl. 10, 16–24.

Gupta, S., Jaafar, J., wan Ahmad, W.F., Bansal, A., 2013. Feature Extraction Using Mfcc. Signal Image Process. An Int. J. 4, 101–108.

Lee, K.F., Hon, H.W., 1989. Speaker-Independent Phone Recognition Using Hidden Markov Models. IEEE Trans. Acoust. 37, 1641–1648.

Muda, L., Begam, M., Elamvazuthi, I., 2010. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv Prepr. arXiv1003.4083.

Nath, S.S., Mishra, G., Kar, J., Chakraborty, S., Dey, N., 2014. A survey of image classification methods and techniques. 2014 Int. Conf. Control. Instrumentation, Commun. Comput. Technol. ICCICCT 2014 554–557.

Sakoe, H., Isotani, R., Yoshida, K., Iso, K. ichi, Watanabe, T., 1989. Speaker-independent word recognition using dynamic programming neural networks. ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 1, 29–32.

Thiang, D.W., 2007. Implementation of speech recognition on MCS51 microcontroller for controlling wheelchair. In: International Conference on Intelligent and Advanced Systems.

Thiang, T., Wijaya, D., 2009. Limited speech recognition for controlling movement of mobile robot implemented on ATmega162 microcontroller. Proc. - 2009 Int. Conf. Comput. Autom. Eng. ICCAE 2009 347–350.

Wijoyo, Suryo, Wijoyo, S, 2011. Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot. In: Proceedings of 2011 International Conference on Information and Electronics Engineering (ICIEE 2011). pp. 28–29.

Yu, D., Deng, L., 2016. AUTOMATIC SPEECH RECOGNITION. Springer.

DOI:https://doi.org/10.26750/Vol(6).no(2).paper 5 http://journal.uor.edu.krd/index.php/JUR Vol.6. No.2, September.2019

پوختە:

سیستەمی ناسینەوەی زمان بابەتێکی بایەخدارە لە پەیوەندی نیوان مرۆڤ و کۆمپیوتەردا ، وەبابەتێکی گرنگە لە لایەن توێژەرەکانەوە ، توانای بە کار خستنی ئامێرە کانە بۆ ناسینە وە ی وشە کان و دە ستە واژە کان لە زمانی ئاخافتنا وە دواتر گۆرینی بۆ شێوازئ کە کۆمپبوتە ر لێ تێ بگات. ناسینە وە ی ژمارە کان بە شیکە لة سیستە می ناسینە وە ي زمان.لەم توێژینەوەیەدا (سێ) کاراکتەری جیاوازی شەپۆلەکانی دەنگ (MFCC,LPC and Format frequency) بەکارهێنراوە بۆ جیاکردنەوەی (دە) ژمارەی زمانی کوردی (0-9) کە ئەم کاراکتەرانە کاراکتەری گشتی شەپۆلەکانن ، دواتر ئەو کاراکتەرانە دەبنە ئەو زانیاریانەی کە ئەلگۆریسمی (Support vector machine) ی پێ مەشق بکرێت بەمەبەستی فێربونی شێوەی تەواوی کاراکتەرەکان بۆ هەریەکە لە ژمارەکان . شێوەی فیرکردنەکە ئەنجام دراوە لەڕێگەی هەریەکە لەکاراکتەرەکان وە پاشان تیکەڵ کردنی کاراکتەراکان . ئەنجامی توێژینەوەکە نیشانی دەدات کەتێکەڵ کردنی کاراکتەرەکان بەشێوازێکی زۆر باش ژمارەکان لەیەکدی جیادەکاتەوە.ژمارە وەک داتا بەکارهاتووە لەم توێژینەوەیە.

خلاصة :

التعرف على الكلام هو موضوع حاسم في مجال التفاعل بين الإنسان والحاسوب. إن قدرة الآلة فى التعرف على الكلمات والعبارات في اللغة المنطوقة هي التعرف على الكلام ثم تحويلها إلى تنسيق يمكن قراءته آليًا. يعد التعرف على الأرقام جزءًا من نظام التعرف على الكلام. في هذا البحث ، تم اقتراح ثلاث ميزات طيفية تشمل MFCC و LPC وترددات التشكيل لتصنيف عشرة ارقام منطوقة باللغة الكردية (0-9). يتم استخراج الميزات بصورة عامة ، وتغذية مصنف SVM ثنائي الاتجاه. تجرى تجارب بما في ذلك كل ميزة فردية وأشكال مختلفة من الانصهار وتظهر النتائج. يعمل دمج الميزات على تحسين النتيجة بشكل ملحوظ ويوضح أن الميزات المختلفة تحمل معلومات تكميلية.. تم اختبار النموذج المقترح على مجموعة البيانات التي تم جمعها في كردستان.

DOI:https://doi.org/10.26750/Vol(6).no(2).paper 5 http://journal.uor.edu.krd/index.php/JUR Vol.6. No.2, September.2019