International Journal of Electrical, Electronic and Communication Sciences
ISSN: 2517-9438
Vol:7, No:12, 2013

Evaluation of Features Extraction Algorithms for a Real-Time Isolated Word Recognition System

Tomyslav Sledevič, Artūras Serackis, Gintautas Tamulevičius, Dalius Navakauskas

Abstract—Paper presents an comparative evaluation of features extraction algorithm for a real-time isolated word recognition system based on FPGA. The Mel-frequency cepstral, linear frequency cepstral, linear predictive and their cepstral coefficients were implemented in hardware/software design. The proposed system was investigated in speaker dependent mode for 100 different Lithuanian words. The robustness of features extraction algorithms was tested recognizing the speech records at different signal to noise rates. The experiments on clean records show highest accuracy for Mel-frequency cepstral and linear frequency cepstral coefficients. For records with 15 dB signal to noise rate the linear predictive cepstral coefficients gives best result. The hard and soft part of the system is clocked on 50 MHz and 100 MHz accordingly. For the classification purpose the pipelined dynamic time warping core was implemented. The proposed word recognition system satisfy the real-time requirements and is suitable for applications in embedded systems.

Keywords—Isolated word recognition, features extraction, MFCC, LFCC, LPCC, LPC, FPGA, DTW.

I. Introduction

SPEECH is the natural and easiest way to communicate between humans. The electronic devices are rapidly evolving and becoming increasingly used for domestic purposes in a few last decades. Therefore, it is desired to interact with computer hardware in the same convenient way as with humans. A lot of researchers solve the challenges related to accuracy and bandwidth improving in existing speech recognition methods in order to particularly facilitate the daily lives with the voice controlled devices.

The main issues, which prevent to design a reliable and universal method for speech recognition, are the environmental impacts, poor pronunciation, speech variability, limited vocabulary size, and a similar phonetic transcription of the words [1]. These issues influence features extraction in a speech. Therefore, an evaluation of algorithms is important especially for the real-time recognizers. The experimental investigation of factors influencing recognition accuracy shows that setting the same training and testing environments yields improved accuracy [2], [3]. Authors claim that features based on cepstrum coefficients are very sensitive to environment. However the auto-regression based features, e.g., linear predictive coefficients (LPC) and linear predictive cepstral coefficients (LPCC), have an average sensitivity. The quality estimation of the speech features shows that linear frequency cepstral coefficients (LFCC), Mel scale cepstral coefficients (MFCC) are suitable for Lithuanian phonemes recognition. Experiments on small vocabulary Lithuanian speech recognition confirmed that accuracy of 93 % is acceptable for application with multilingual transcriptions engines [4].

In a last few years the MFCC and LPCC features become a reasonable leaders in the recognition systems. A 98.5 % rate was achieved recognizing isolated words in a small vocabulary using MFCC [5]. The MFCC and LPCC features are suitable for classification of speech disfluencies with 92.55 % and 94.51 % rates accordingly [6]. A comparative study of LFCC vs MFCC was performed in [7]. Results show that LFCC consistently outperforms MFCC. The benefits are visible especially on female speech recognition. There are known hardware-based MFCC and LPCC implementations, which allows to accelerate features extraction process [8], [9]. The combination of MFCC and LPCC are also appropriate for speaker identification with maximum 97.12 % [10]. The LPCC always outperforms the LPC features over normal and noisy conditions [11]. The highest 91.4 % recognition rate was achieved using LPC and artificial neural network in a small vocabulary system [12]. The auto-regression based algorithm (e.g., LPC) are more suitable for software implementation rather than hardware due to precision requirements [13]. Therefore, a soft-core processor is popular for recursion implementation [14].

The LFCC features extraction algorithm was implemented fully in hardware and presented in our previous work [15]. It gives an average 97 % isolated word recognition accuracy using six sessions for each speaker in order to create a dictionary. The size of dictionary plays a significant role particularly in embedded voice controlled systems. Large dictionary allows to minimize recognition error while small dictionary speeds-up the classification process. In this paper the MFCC, LFCC, LPCC and LPC features extraction algorithms are evaluated at different signal to noise rates with the aim to determine robust features for proposed FPGA-based isolated word recognition system.

Even if features extraction algorithm works correct, there is still important to chose proper classification method. The dynamic time warping (DTW) method is an appropriate way to find similarities in two vectors of features. The Field Programmable Gate Array (FPGA) based pipelined implementation of the classification allows to accelerate speech recognition process [16], [17]. A comparative study of DTW implementations on different platforms shows that FPGA-based DTW outperforms the GPU and CPU-based implementations more than 44 and 2100 times

T. Sledevič, A. Serackis, G. Tamulevičius and D. Navakauskas are with the Department of Electronic Systems, Vilnius Gediminas Technical University, Naugarduko St. 41-413, 03227 Lithuania (e-mail: tomyslav.sledevic@vgtu.lt).

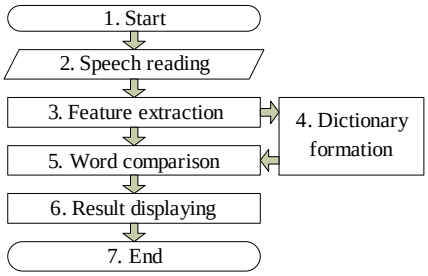

Fig. 1 Isolated word recognition algorithm

accordingly [18], [19]. These arguments inspires to implement DTW intellectual property (IP) core for a real-time isolated word recognition.

In the Section II the isolated word recognition system is presented. In Section III the implemented and investigated features extraction algorithms are described. In Section IV the results on isolated word recognition experiments are summarized. General conclusions are stated in Section V.

II. Isolated Word Recognition System

Whole recognition system is implemented in single FPGA chip. The word recognition algorithm consists of several steps, as shown in Fig. 1. The speech signal can be loaded from SD memory card or microphone. The extracted features are stored in the dictionary. Single IP cores are used for all features extraction and comparison algorithms. Each features extraction algorithms process 256 samples 8 bit 11.025 kHz speech data. There are reserved 2^{14} samples for each isolated word regardless of its real length. Each word has 128 features vectors. Therefore, the size of DTW matrix is always permanent and equal to 128×128. The hardware part of the proposed system is clocked at 50 MHz while software part – at 100 MHz.

There are possible real-time and non-real-time recognizing modes in the proposed system. In the first mode, the speech signal is constantly captured from microphone. In the second, signal is read from external memory. An important for the real-time is to satisfy the maximal delay requirement of 11.61 ms as a time between incoming frames. Each implemented IP core meets this condition. Soft-core processor-based LPCC and MFCC calculations have a highest delay of 3.35 ms comparing to all others cores. The 11.61 ms period is enough to compare 90 isolated words with single DTW core [15].

The block diagram in Fig. 2 shows the connectivity between IP cores with bus width information. Each IP core in the proposed system works independent and communicate with others via synchronization signals. The soft-core processor was implemented for the IP cores that require precision and operation on floating point numbers. The decision maker is a linear function that sequentially compares back-path errors given from DTW IP core. In the real-time mode system firstly looks for activation word. If this word is recognized then second spoken command is compared with dictionary in a

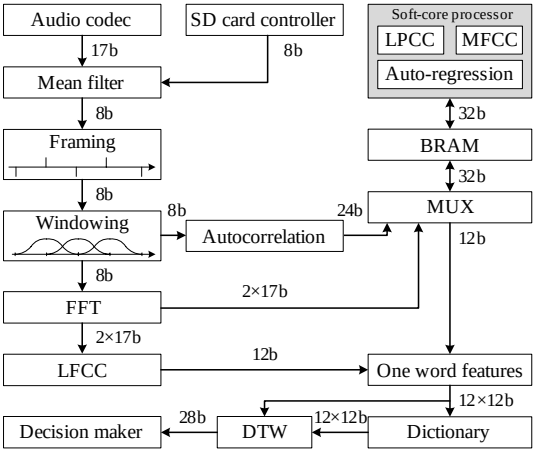

Fig. 2 The block diagram of the isolated word recognition system

Require: Store all features vectors in BRAM
for all tested records **do**
  for all words stored in dictionary **do**
    1. Compute first element $e(0,0)$ in error matrix:
    $d_{\mathrm{e}} = \sqrt{\sum_{i=1}^{12} (c_{\mathrm{w}}(i) - c_{\mathrm{d}}(i))}$;
    2. Compute first line in error matrix:
    for x in 1 to 127 **do**
      $e(x,0) = d_{\mathrm{e}} + e(x-1,0)$;
    end for
    3. Compute first column in error matrix:
    for y in 1 to 127 **do**
      $e(0,y) = d_{\mathrm{e}} + e(0,y-1)$;
    end for
    4. Fill the rest of the area in error matrix:
    for x in 2 to 127 **do**
      for y in 2 to 127 **do**
        $e(x,y) = d_{\mathrm{e}} + \min(e_1, e_2, e_3)$;
      end for
    end for
    5. Find the back-path error $E_{\mathrm{back}}^{\mathrm{present}}$.
    if $E_{\mathrm{back}}^{\mathrm{present}} < E_{\mathrm{back}}^{\mathrm{previous}}$ **then**
      Update the *index* of recognized word.
    end if
  end for
end for
return vector of *indexes* for all tested records

Fig. 3 Pipelined dynamic time warping implementation

features space. The implementation of comparison process is presented in algorithm on Fig. 3.

At the beginning of DTW process the features vectors of the spoken word must be stored in one word features memory. The hardware-based non-restoring square root algorithm was used to evaluate the Euclid distance d_{e} between features. The distance calculation have a pipelined structure. The 2×144 bit features come to the DTW IP core at each rising edge of clock. The features matching errors are stored in $128 \times 128 \times 20$ bit size error matrix based on BRAM. The address of error matrix memory, address of features memory and d_{e} must be calculated synchronously while calculating the steps 2–4 in algorithm (Fig. 3). Because the calculation of d_{e} brings additional delay, therefore the address counters

Require: Prepare a 128 samples of speech signal x.
1. 1. Apply the Hann window function: $s = x \cdot h$.
2. 2. Chose desired features extraction algorithm.
3. **if** MFCC or LFCC **then**
4. Calculate spectrum $S = \mathrm{FFT}(s)$.
5. **if** LFCC **then**
6. $C = \mathrm{real}\left(\mathrm{FFT}\left(\log_2\left(\mathrm{abs}(S)\right)\right)\right)$.
7. **else**
8. A. Save spectrum in BRAM and start soft-core.
9. B. Apply Mel-scale filter banks, $\log_2$ and DCT.
10. C. Return MFCC coefficients to hardware part.
11. **end if**
12. **else**
13. A. Calculate 13 first autocorrelation coefficients.
14. B. Save coefficients in BRAM and start soft-core.
15. C. Apply Levinson-Durbin recursion.
16. **if** LPC **then**
17. Return LPC coefficients to hardware part.
18. **else**
19. Recalculate $c_i = a_i + \sum_{j=0}^{i-1} j c_j a_{i-j}/i$.
20. Return LPCC coefficients to hardware part.
21. **end if**
22. **end if**
23. **return** vector of features.

Fig. 4 Features extraction for one frame speech signal

are implemented in separate processes. The error $e(x, y)$ and back-path error E_{back} are estimated using sliding 2×2 window [15]. The DTW IP core returns an vector of indexes corresponding recognized records. Two words comparison takes 6404 clock pulses using DTW with border constrains.

III. Investigated Features Extraction Algorithms

Features are extracted from frames contained 256 samples of speech. The implementation of features extraction process is presented in algorithm shown on Fig. 4. The LFCC algorithm is implemented fully in hardware. It uses Radix-2 based FFT IP core returning the spectrum after 1667 clock pulses. Simplified log_2 IP core searches the highest bit in the binary number. The LFCCs are calculated in 66.4 μs. The MFCC, LPC and LPCC has partial hardware/software implementation. The 20 Mel-scale filter banks gives best recognition result for tested data set. The coefficients of the filters are pre-calculated and stored in look-up table. The log_2 and discrete cosine transform (DCT) are implemented in soft-core. First MFCC is ignored and the next 12 coefficients form features vector. The calculation of MFCC takes 3304.5 μs.

The 13-th order autocorrelation has pipelined hardware implementation and takes 5.4 μs. The Levinson-Durbin recursion is software-based and is computed in 1662.1 μs. The recalculation of LPCC takes 1682.0 μs. Dual port BRAM is used to return the features vector form software to hardware part of system. When the last 128th features vector is stored then the enable signal is sent to the DTW core.

IV. Isolated Word Recognition Experiments

The dictionary contains the isolated Lithuanian words pronounced by 5 male and 5 female speakers. 4 sessions of 100 words were recorded for each speaker in the office environment. The first session was used for the training and the rest 3 sessions were used for the testing. The recognition rate was tested using original records and using the same words with 30 dB and 15 dB SNR. The aim is to determine most suitable features extraction algorithm among MFCC, LFCC, LPCC and LPC for speaker dependent recognizer. The averaged results over male and female speakers are given in Table I. The best average recognition rate was achieved for original records using MFCC (93.2%) and LFCC (93.0%) features. The LFCC (90.8%) and little less LPCC (89.8%) features are more robust to noise when SNR is 30 dB. The LPCC (83.7%) features give better recognition accuracy when SNR is 15 dB. The LPC features give the worst recognition rates in all cases, comparing with others features extraction algorithms.

TABLE I
Averaged Isolated Word Recognition Results

Features	Original records	Records with SNR = 30 dB	Records with SNR = 15 dB
MFCC	93.2	86.7	79.3
LFCC	93.0	90.8	72.4
LPCC	91.5	89.8	83.7
LPC	82.5	78.6	67.1

TABLE II
Isolated word recognition rates for male speakers (%)

Speaker	M_1	M_2	M_3	M_4	M_5	Average
MFCC						
Original	100	98	91	83	86	91.6
30 dB	90	95	83	76	78	84.4
15 dB	87	92	77	62	66	76.8
LFCC						
Original	99	97	91	84	89	92.0
30 dB	96	94	79	83	91	88.6
15 dB	91	72	65	69	75	74.4
LPCC						
Original	99	95	90	77	79	88.0
30 dB	94	91	77	84	88	86.8
15 dB	93	85	72	79	77	81.2
LPC						
Original	91	85	73	68	57	74.8
30 dB	79	82	57	66	64	69.6
15 dB	82	64	48	57	56	61.4

The recognition accuracy highly depends on speaker gender, tone of voice, articulation and speed of pronunciation. It is important to diagnose the reasons of differences in rates while using same features extraction method. Each entry in the Table II and Table III evaluates the recognition accuracy over three times repeated experiments. In the male (M_i) speaker case there are two speakers (M_1 and M_2) with high recognition accuracy over all features extraction methods and different SNRs. There are speakers (M_4 and M_5) with relatively lower rates comparing with others. These two speakers have low tone of voice, therefore the spectral information is distributed in low frequency domain. Some errors are caused by different pronunciation of the same word in the following sessions.

In the case of female speakers the F_2 has lowest rate. In the F_2 records the low-frequency periodic noise is observed.

TABLE III
ISOLATED WORD RECOGNITION RATES FOR FEMALE SPEAKERS (%)

Speaker	F_1	F_2	F_3	F_4	F_5	Average
MFCC						
Original	98	87	98	99	92	94.8
30 dB	94	82	91	97	81	89.0
15 dB	92	76	83	93	65	81.8
LFCC						
Original	95	86	95	99	95	94.0
30 dB	95	85	93	98	94	93.0
15 dB	67	65	72	81	67	70.4
LPCC						
Original	97	89	98	99	92	95.0
30 dB	97	85	93	99	90	92.8
15 dB	87	74	83	98	89	86.2
LPC						
Original	93	87	92	97	82	90.2
30 dB	89	81	92	95	81	87.6
15 dB	72	66	73	87	66	72.8

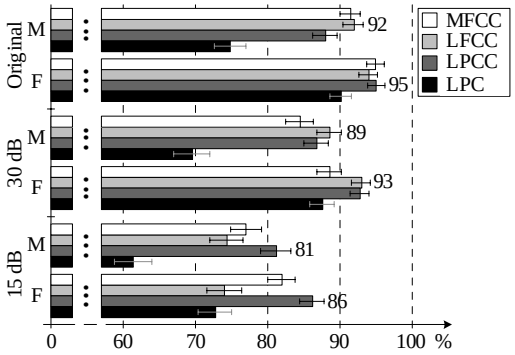

Fig. 5 The average recognition rate of isolated word pronounced by male and female speaker using different features extraction algorithms

The speaker F_5 pronounces the words faster than F_1, F_3 and F_4. It influences the accuracy, as shown in Table III.

The average recognition rate of the female records outperforms the rate of male records in all features extraction cases (Fig. 5). Other researchers declare similar 91–96 % recognition rates in the speech recognition systems based on software [4], [7], [12] and hardware [16], [17], [20] implementations.

V. CONCLUSION

The eligibility of MFCC, LFCC, LPCC and LPC features extraction algorithm was evaluated for real-time isolated word recognition system based on FPGA. The results of the experiments on original records show that LFCC and MFCC features have approximately equal robustness. The application of MFCC and LPCC is more suitable for recognition of clean records. The LFCC and LPCC features are appropriate for 30 dB environment noise. The LPCC features gives the best accuracy at 15 dB SNR comparing to other algorithms at same noise level. The recognition rate is higher for female records over all SNR and features extraction algorithms. It is issued by distribution of the spectral components in low as well as in higher frequency domain.

ACKNOWLEDGMENT

This research was funded by a grant (No. MIP-092/2012) from the Research Council of Lithuania.

REFERENCES

[1] G. Tamulevičius, "Isolated word recognition systems implementation," Ph.D. dissertation, Vilnius Gedinimas Technical University, Vilnius, May 2008. [Online]. Available: http://www.mii.lt/files/mii_dis_08_tamulevicius.pdf
[2] G. Čeidaitė and L. Telksnys, "Analysis of Factors Influencing Accuracy of Speech Recognition," *Electronics and Electrical Engineering*, vol. 105, no. 9, pp. 69–72, Nov. 2010.
[3] R. Leleikytė and L. Telksnys, "Quality Estimation Methodology of Speech Recognition Features," *Electronics and Electrical Engineering*, vol. 110, pp. 113–116, May 2011.
[4] R. Maskeliūnas and A. Esposito, "Multilingual Italian-Lithuanian Small Vocabulary Speech Recognition via Selection of Phonetic Transcriptions," *Electronics and Electrical Engineering*, vol. 121, pp. 85–88, Jun. 2012.
[5] K. A. Darabkh, A. F. Khalifeh, B. A. Bathech, and S. W. Sabah, "Efficient DTW-Based Speech Recognition System for Isolated Words of Arabic Language," *World Academy of Science, Engineering and Technology*, vol. 77, pp. 85–88, May 2012.
[6] O. C. Ai, M. Hariharan, S. Yaacob, and L. S. Chee, "Classification of Speech Dysfluencies with MFCC and LPCC Features," *Expert Systems with Applications*, vol. 39, no. 2, pp. 2157–2165, Feb. 2012.
[7] X. Zhou, D. Garcia-Romero, R. Duraiswami, C. Espy-Wilson, and S. Shamma, "Linear versus Mel Frequency Cepstral Coefficients for Speaker Recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, Dec. 2011, pp. 559–564.
[8] N. V. Vu, J. Whittington, H. Ye, and J. Devlin, "Implementation of the MFCC Front-End for Low-Cost Speech Recognition Systems," in *Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS 2010)*, Jun. 2010, pp. 2334–2337.
[9] M. Staworko and M. Rawski, "FPGA Implementation of Feature Extraction Algorithm for Speaker Verification," in *Proceedings of the 17th International Conference on Mixed Design of Integrated Circuits and Systems (MIXDES 2010)*, Jun. 2010, pp. 557–561.
[10] Y. Yujin, Z. Peihua, and Z. Qun, "Research of speaker recognition based on combination of LPCC and MFCC," in *IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS 2010)*, vol. 3, Oct. 2010, pp. 765–767.
[11] C. Y. Fook, M. Hariharan, S. Yaacob, and A. Ah, "Malay speech recognition in normal and noise condition," in *IEEE 8th International Colloquium on Signal Processing and its Applications (CSPA 2012)*, Mar. 2012, pp. 409–412.
[12] T. Wijoyo and S. Wijoyo, "Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot," in *International Conference on Information and Electronics Engineering IPCSIT*, vol. 6, 2011, pp. 179–183.
[13] J. Xu, A. Ariyaeeinia, and R. Sotudeh, "Migrate Levinson-Durbin based Linear Predictive Coding algorithm into FPGAs," in *12th IEEE International Conference on Electronics, Circuits and Systems (ICECS 2005)*, Dec. 2005, pp. 1–4.
[14] M. Atri, F. Sayadi, W. Elhamzi, and R. Tourki, "Efficient Hardware/Software Implementation of LPC Algorithm in Speech Coding Applications," *Journal of Signal and Information Processing*, vol. 3, no. 9, pp. 122–129, 2012.
[15] T. Sledevic and D. Navakauskas, "FPGA-Based Fast Lithuanian Isolated Word Recognition System," in *EUROCON, 2013 IEEE*, Jul. 2013, pp. 1630–1636.
[16] G. Zhang, J. Yin, Q. Liu, and C. Yang, "A Real-Time Speech Recognition System Based on the Implementation of FPGA," in *Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC 2011)*, vol. 2, Jul 2011, pp. 1375–1378.
[17] S. T. Pan and X. Y. Li, "An FPGA-Based Embedded Robust Speech Recognition System Designed by Combining Empirical Mode Decomposition and a Genetic Algorithm," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, pp. 2560–2572, Sep 2012.
[18] D. Sart, A. Mueen, W. Najjar, E. Keogh, and V. Niennattrakul, "Accelerating Dynamic Time Warping Subsequence Search with GPUs and FPGAs," in *IEEE 10th International Conference on Data Mining (ICDM 2010)*, Dec 2010, pp. 1001–1006.

[19] Y. Zhang, K. Adl, and J. Glass, "Fast spoken query detection using lower-bound Dynamic Time Warping on Graphical Processing Units," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, Sep 2012, pp. 5173–5176.
[20] O. Cheng, W. Abdulla, and Z. Salcic, "HardwareSoftware Codesign of Automatic Speech Recognition System for Embedded Real-Time Applications," *IEEE Transactions on Industrial Electronics*, vol. 58, 2011.