ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Implementation on Gender-Driven Emotion Recognition through Speech Signals for Ambient Intelligence Applications

Rushikesh Gade[1] , S.R.Gulhane[2]

PG Student, Department of E&TC, D.Y.Patil College of Engineering, Talegaon Savitribai Phule University of Pune, Pune, India

Professor, Department of E&TC, D.Y.Patil College of Engineering, Talegaon Savitribai Phule University of Pune, Pune, India

ABSTRACT: This brief presents an energy-efficient architecture to extract mel-frequency cepstrum coefficients (MFCCs) for real-time speech recognition systems. Based on the algorithmic property of MFCC feature extraction, the architecture is designed with floating-point arithmetic units to cover a wide dynamic range with a small bit-width. Moreover, various operations required in the MFCC extraction are examined to optimize operational bit-width and lookup tables needed to compute nonlinear functions, such as trigonometric and logarithmic functions. In addition, the dataflow of MFCC extraction is tailored to minimize the computation time. As a result, the energy consumption is considerably reduced compared with previous MFCC extraction systems. The obtained results show also that the features selection adoption assures a satisfying recognition rate and allows diminishing the employed features. Future improvements of the proposed solution may include the implementation of this system over mobile devices such as smartphones.

KEYWORDS- Emotional speech recognition, MFCC(mel-frequency cepstrum coefficients), SVM(Support Vector Machine).

I.INTRODUCTION

speech recognition has widely been used in the last decade, and its importance becomes higher as the era of the Internet of Things comes close to reality. Due to the prevalence of energy-limited devices, energy-efficient architecture is inevitably demanded to lengthen the device life. The demand for low-energy architecture leads to the speech recognition system being implemented with dedicated hardware units. A speech recognition system consists of two processes:

1) feature extraction and
2) classification.

The feature extraction process picks the characteristics of a sound frame, and a word is selected in the classification process by analyzing the extracted features. This brief mainly focuses on the hardware design of feature extraction. The most widely known feature extraction is based on the mel-frequency cepstrum coefficients (MFCCs), as MFCC-based systems are usually associated with high recognition accuracy [5]. MFCC extraction was implemented with an optimized recognition program running on a low-power reduced instruction set computer processor platform. To reduce energy consumption further, dedicated architectures have been proposed and constructed with fixed-point operations. The previous architectures, however, have not fully considered the arithmetic property of the MFCC extraction algorithm. This brief presents a new energy-efficient architecture for MFCC extraction. Investigating the algorithmic property of MFCC extraction, we renovate the previous architecture with optimization techniques to reduce both hardware complexity and computation time. As a result, the energy consumption is remarkably reduced compared with the previous architectures.

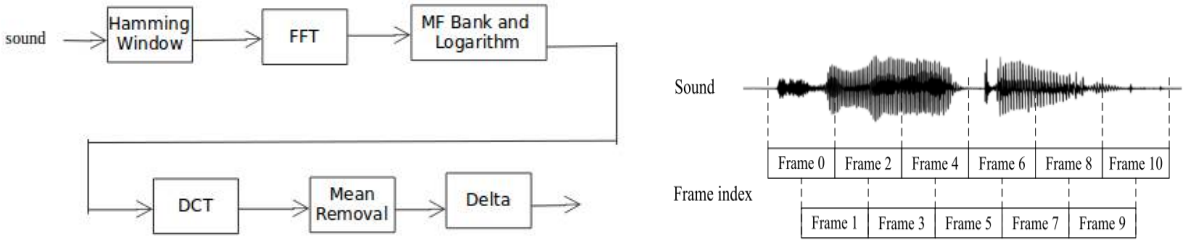
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017



(a) (b)

Fig.1 (a) Modified MFCC Extraction System (b) Sound Frame

This concise presents a new energy-efficient architecture for MFCC extraction. Fig1 describe the modified MFCC extraction system and the improvement systems. This concise presents a new energy-efficient architecture for MFCC extraction. Exploring the algorithmic property of MFCC extraction, we renovate the previous architecture with improvement strategies to reduce both hardware complexity and computation time. Subsequently, the energy consumption is remarkably reduced compared with the previous architectures.

II.LITERATURE SURVEY

This brief presents an energy-efficient architecture to extract mel-frequency cepstrum coefficients (MFCCs) for real-time speech recognition systems. Based on the algorithmic property of MFCC feature extraction, the architecture is designed with floating-point arithmetic units to cover a wide dynamic range with a small bit-width[1].They propose a new Fourier parameter model using the perceptual content of voice quality and the first- and second-order differences for speaker-independent speech emotion recognition. Experimental results show that the proposed Fourier parameter (FP) features are effective in identifying various emotional states in speech signals. They improve the recognition rates over the methods using Mel frequency cepstral coefficient (MFCC) features[2]. This paper proposes a system that allows recognizing a person's emotional state starting from audio signal registrations. The system is composed of two subsystems: 1) gender recognition (GR) and 2) emotion recognition(ER). The experimental analysis shows the performance in terms of accuracy of the proposed ER system. The results highlight that the a priori knowledge of the speaker's gender allows a performance increase [3].

A robust speech emotion recognition system relies on a large number of training data, which are difficult to collect in practice. To tackle this problem, a novel speech emotion recognition method based on hidden factor analysis is presented. By utilising the mixture of factor analysers approach, the acoustic features are decomposed into an emotion-independent component and an emotion-specific component. The emotion-specific component, described by a low-dimensional emotion identity vector, is adopted for classification. The proposed approach is evaluated via cross-corpus emotion recognition, and the experimental results demonstrate the efficacy of the proposed method. Kotti and Paternò [5] extracted 2,327 features in total for speaker-independent recognition that were related to the statistics of pitch, formants, and energy contours as well as spectrum, cepstrum, autocorrelation, voice quality, jitter, shimmer and others.

III.FEATURES EXTRACTION

A) MFCC Extraction Features

MFCC was initially introduced and applied to speech recognition in [16]. It has been prominently utilized for speech emotion recognition [17]. By considering the reaction of human ears to various frequencies, the Mel frequency is determined according to the characteristics of human audition. In this study, MFCC features were extracted for comparison with the proposed FP features. For emotion recognition, MFCC features usually include mean, maximum, minimum, median, and standard deviation. All speech signals were initially filtered by a high-pass filter with a pre-emphasis coefficient of 0.97. The first 13 MFCCs and the associated delta- and double-delta MFCCs were extracted to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 6, June 2017

form a 39-dimensional feature vector. Its mean, maximum, minimum, median and standard deviation were further derived out, which led to a 195-dimensional MFCC feature vector in total.

MFCC extraction is a procedure that extracts features representing the characteristics of a sound frame. As appeared in Fig. 1 (b) sound frame consists of N sound samples and it is half-overlapped in the time domain with the previous and next frames. The overall flow of the conventional MFCC extraction is shown in Fig. 1 (a), which extracts MFCC vectors for a sound frame s = {s[0], s[1], . . . , s[N − 1]}. The MFCC vectors consist of {C, C ‘, C “ }, where the first feature vector, C = {C[0], C[1], . . . , C[M − 1]}, is an arrangement of (M − 1) MFCCs and the logarithmic energy of the sound signals contained in the frame, and C ‘ and C “ indicate the first and second derivatives of C, respectively. The rest of this section explains how MFCC vectors are computed in detail. The given signal s[n] is pre-processed by applying pre-emphasis and Hamming windowing sequentially. The underlining concept of the pre-emphasis is to amplify high-frequency components obtained by passing through a high-pass filter. The filtered output is given by

$$p[n] = s[n] - 0.97 \cdot s[n-1]$$

Afterward, it is required to degrade p[n] by multiplying the following Hamming window function in order to compensate the overlap between the neighbouring frames. The degraded signal is
computed as

$$h[n] = p[n] \{ 0.54 - 0.46 \cdot \cos(2\pi n/(N-1)) \}$$

After the preemphasis and hamming windowing are completed, the logarithmic energy of the sound frame is calculated as

$$C[0] = \log (\sum h[n])$$

where C[0] is the first component of C.

Since many operations used in the MFCC algorithm depend on complex functions, for example square and logarithmic functions, their outputs are associated with a large dynamic range. Compared with the fixed-point representation, the floating-point representation can cover such a large dynamic range with a much smaller number of bits. In addition, the operation bit-width can be reduced further, grounded on the property that the resulting feature vectors are influenced by the order of magnitude of interim values. For these reasons, a floating-point representation is employed in this brief to implement the modified MFCC extraction algorithm described above. The floating-point number system has a estimation value of

$(-1)^{S} x F x 2^{E}$

where S, E, and F denote the sign, exponent, and fraction parts, respectively. The bit-width of E is set to a constant value larger than or equal to the upper bound in order to prevent any possible overflows and underflows, but the bit-width of F is determined by conducting two optimization processes.

1) Modified MFCC Extraction Features

This segment introduces a new floating-point MFCC extraction architecture derived to understand the MFCC extraction with a small hardware cost. This approach is totally unique from that have used a different hardware unit for each operation. The proposed architecture is described with setting N to 256, M to 13, and L to 32. For sound signals sampled with 16 bits at 16 kHz, in addition, the bit-widths of F and E in the floating-point representation are determined to 6 and 7 bits, respectively. In speech recognition system, One of the buffers stores a half of a sound frame and the other buffer is used to save the remaining data of the frame. Since subsequent frames share a half frame, only one buffer is updated for the next sound frame.

By analyzing the dataflow of the modified MFCC algorithm, we propose a new MFCC extraction system implementable with a small hardware cost. In the general design or architecture of the proposed system consists of a multiply-and-accumulate (MAC) unit, an address generation unit, a controller, memories, and counters.

The output of the MAC unit are saved into one of four memories: 1) general purpose registers (GPRs); 2) register files (Rfs); 3) C buffers (CBs); and 4) C’ buffers (CBFs). The GPRs are utilized to store intermediate values such as the interim sound energy of a frame. The RFs are included to successfully compute such processes storing many values as FFT, mel filtering, DCT, and derivative computations. Grounded on the dataflow analysis of the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering
(An ISO 3297: 2007 Certified Organization)
Website: www.ijircce.com
Vol. 5, Issue 6, June 2017

modified MFCC algorithm, an efficient memory structure consisting of four separate RFs is determined. The proposed architecture utilizes two counters to generate two addresses needed to access two memories simultaneously. The controller controls all the blocks to process the MFCC extraction algorithm, and its main role is to decide the input signals to be fed to the MAC unit and the storage to be used to store the results.

B) Pitch Pattern Feature

The prosody of synthetic speech is generally not the same as natural speech and therefore the pitch pattern is another good candidate feature for a countermeasure. The pitch pattern, Φ[n,m], is calculated by dividing the short-range autocorrelation function, r[n, m] by a normalization function, p[n, m] which is proportional to the frame energy

$$\Phi[n, m] = \frac{r[n,m]}{p[n,m]}$$

where

$$r[n, m] = \sum_{k=-m/2}^{m/2} x\,[n + k - m/2]x[n + k + m/2]$$

$$p[n, m] = \frac{1}{2}\sum_{k=-m/2}^{m/2} x^2\,[n + k - m/2] + \frac{1}{2}\sum_{k=-m/2}^{m/2} x^2\,[n + k + m/2]$$

and n, m are the sample instant and lag, respectively, over which the autocorrelation is computed. The lag parameter is chosen such that pitch frequencies can be observed [72]; in this work, we choose $32 \leq m \leq 320$ for a sample rate of 16kHz. Once the pitch pattern is computed, we segment it into a binary pitch pattern image through the rule

$$\Phi seg\ [n,m] = \sum 1\ \Phi[n,m] \geq \Theta;\ 0\ \Phi[n,m] < \Theta$$

where θ is a threshold; we set θ = 1/ 2 for all n, based on preliminary results on the development set. An example pitch pattern image is shown in Fig. 2.

Extracting features from the pitch pattern is a two-step process: 1) computation of the pitch pattern; 2) image analysis. First, the pitch pattern is computed using (4) and segmented using to form a binary image. In the second step, image processing of the segmented binary pitch pattern is performed in order to extract the connected components (CCs), i.e., black regions in Fig. 2. This processing includes determining the bounding box and area of a CC, which are then used to distinguish between two types of CC: pitch pattern connected components (PPCC) and irregularly-shaped components or artefacts.

The resulting CCs are then analysed and the mean pitch stability μ s , mean pitch stability range μ R , and time support (TS) of each CC are computed. The proposed image processing-based approach determines parameters on a per connected component basis and then computes statistics over the connected components of the utterance. The six element utterance feature vector used for classification contains μ R and the TS of the artefacts, the number of artefacts, μ S and T S of the PPCC, and standard deviation of the TS of PPCC. Other utterance features were considered during the training and development stage but were found not to contribute to the classifier accuracy.

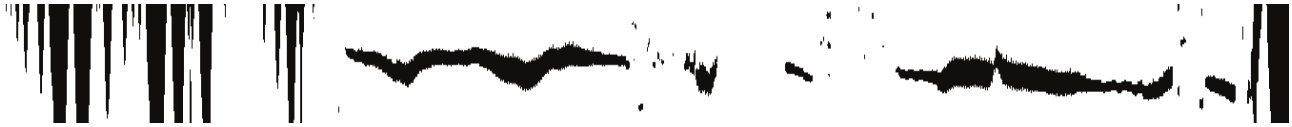


Fig. 2 Example binary pitch pattern image illustrating pitch stability S c, pitch stability range R c , upper edge τ U , lower edge τ L , connected component time support, and artefacts.

C) Formant

The parameters related to vocal tract is formant frequency or formants. When the excitation passes through vocal tract it resonates at certain resonating frequency. These resonating frequencies are termed as formants [6]. Formant can be calculated with linear cepstral prediction (LPC). LPC is speech analysis technique used for establishing pitch, formants, vocal tract functions. The working behind LPC is to predict the present sample from the previous sample. The objective is to find coefficients and G. In source filter model, filter is constrained to be pole linear filter.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Current sample = (past sample * coefficient) + excitation
S[n] equation gives linear combination of past p number of samples.

D) Tone and Gender Recognition Algorithm

The proposed GR method is designed to distinguish a male from a female speaker and has been thought to be realized over mobile devices, such as smart phones. The chosen feature is the mean of the Probability Density Function (PDF), of a number of frames of the voice signal, as explained below.

The signal to be classified as ‘‘Male’’ or ‘‘Female’’ is identified as s(n), n = 1 . . . N . The GR method introduced in this paper is composed of the following steps

1) The signal s(n) is divided into frames.
2) The pitch frequency for each frame is estimated.
3) A number of frames of s(n) is grouped into a odd-number of blocks.
4) The pitch PDF is estimated for each block.
5) The mean of each pitch PDF (PDF mean) is computed.
6) The decision about ‘‘Male’’ or ‘‘Female’’ is taken, for each block, by comparing their PDF mean with a fixed threshold γ thr computed by using the training set.
7) The final decision on the whole signal gender is taken by the majority rule: the signal s(n) is classified as ‘‘Male’’ if the majority of its blocks are classified as ‘‘Male’’. Otherwise, it is classified as ‘‘Female’’. Average pitch frequencies for male and female speakers and the individuated threshold γthr , referred to a recording of 20 blocks are shown in Fig.3

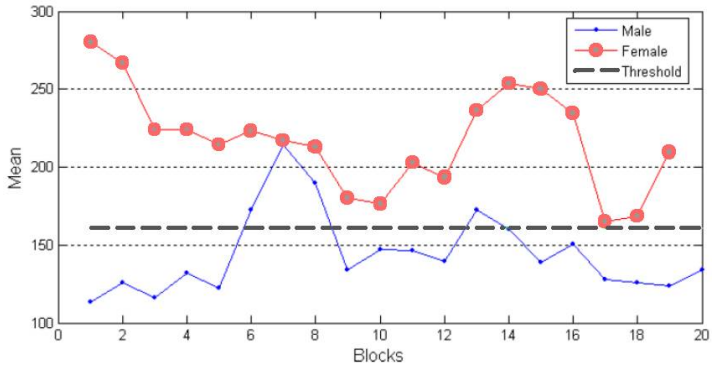


Fig.3 Average pitch frequencies, referred to male and female single speakers, and the employed threshold γ thr , for a recording divided in 20 blocks.

E) Classifier

1. Support Vector Machine

SVM is a binary classifier to analyse the data and recognize the pattern for classification. The main goal is to design hyper plane that classifies all the training vectors in different classes. The aim is to determine functions which obtain the hyper plane. Hyper plane separates two classes of data sets. The linear classifier is defined as the optimal separating hyper plane.

The data sets can be separated in two ways: linearly separated or nonlinearly separated. The vectors are said to be optimally separated if they are separated without error and the distance between the two closest vector points is maximum.

SVM is to find the optimal separating hyper plane which separates two different label sets. Given a set of data $\{x_i, y_i\}$, i =1..n where$x_i \in R^d$ denotes the input vector, $y_i \in \{+1, -1\}$ denotes the output value. Optimal separating hyper plane formula is as:

$w \cdot \Phi(x) + b = 0$

where x is input vector, w is weight vector, and Φ () is a mapping function in non-linear SVM. When linear SVM cannot solve the problem, non-linear SVM uses kernel function to project vector to higher dimensional space [16].

Then SVM finds a linear separation linear hyper plane from high dimensions. SVM can be formulated as following Optimal problem:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

Minimize $\Phi(w) = \frac{1}{2}\|w\|$

Subject to $y_i\ (w \cdot \Phi\ (x_i) + b) \geq 1$

Above the optimal question have solution if and only if it exists one optimal separating hyperplane to separate data perfectly. For non-separable case, it must add a slack variable ξ to release the restrict condition. Then, new optimal problem is:

Minimize

$\Phi(w, \xi i) = \frac{1}{2}\|w\| + c \sum_{i-1}^{N} \xi i$

Subject to

$y_i(w \cdot \Phi(x) + b) \geq 1 - \xi i, \xi i \geq 0\ \forall i$

where C is the penalty parameter of the error term. The decision function of SVM is defined as:

$f(x) = \text{sign}(\sum_{i=1}^{N} \alpha i y_i k(x, x_i) + b)$

where α i is Lagrange multipliers and k () is kernel function.

F) Database

Earlier the database used for automatic speech emotion recognition research was based on acted speech and the researcher tends to have real based data.

Database is divided into three types.

1. Acted speech: actors are asked to express deliberately the human emotions which are predefined.
2. Real life speech: the natural response to the conversion of human with spontaneous reaction which are authentic in nature. Example: call center.
3. Elicited emotional speech in which the emotions are induced with self-report instead of labelling, where emotions are provoked and self-report is used for labelling control.

IV.RESULT

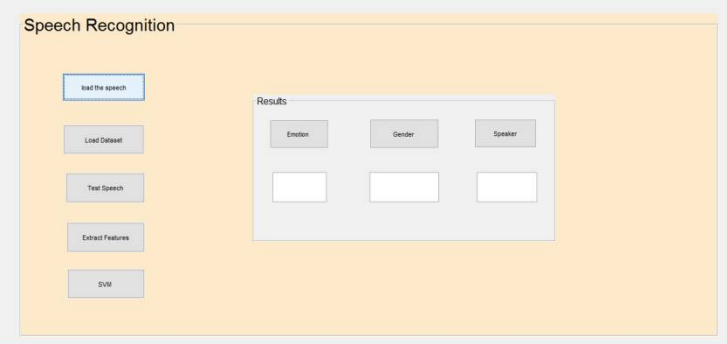


Fig.4 Speech Recognition

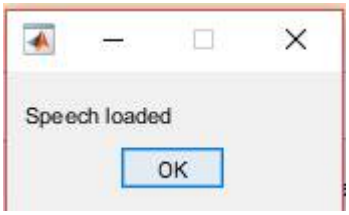


Fig.5 Speech load

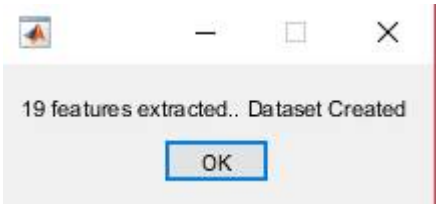


Fig.6 Feature Extraction

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 6, June 2017

V.CONCLUSION

An energy-efficient MFCC extraction architecture has been presented for speech recognition. The MFCC extraction algorithm is modified to minimize computation time without degrading the recognition accuracy noticeably. In addition, the proposed architecture employs floating-point arithmetic operations to minimize the operation bit-width and the total size of LUTs. Furthermore, a floating-point MAC unit and memories are shared with many processes to reduce hardware complexity and energy consumption remarkably. The effectiveness of energy consumption makes the proposed architecture a promising solution for energy-limited speech recognition systems.

REFERENCES

[1]Jihyuck Jo, Hoyoung Yoo, and In-Cheol Park, "Energy-Efficient Floating-Point MFCC Extraction Architecture for Speech Recognition Systems", IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 24, Issue :2, feb 2016.
[2]kunxia wang, ning an, bing nan li,yanyong zhang and lian li, "speech emotion recognition using fourier parameters", ieee transactions on affective computing, Vol. 6, no. 1, january-march 2015.
[3] Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese, and Andrea Sciarrone, "Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications", IEEE Transactions on Emerging Topics in Computing, Vol. 1, Issue: 2, Dec.2013.
[4]Peng Song, Yun Jin, Cheng Zha and Li Zhao, "Speech emotion recognition method based on hidden factor analysis", Electronics Letters, Vol. 51, Issue: 1, january 2015.
[5] M. Kotti and F. Paternó, "Speaker-independent emotion recognition exploiting a psychologically-inspired binary cascade classification schema," Int. J. Speech Technol., vol. 15, pp. 131–150, 2012.