Objective Gender and Age Recognition from Speech Sentences

Fatima K. Faek

Electrical Department, Engineering College, Salahaddin University
Zanko Street, Kirkuk road, Erbil, Kurdistan Region of F.R. Iraq

***Abstract*—In this work, an automatic gender and age recognizer from speech is investigated. The relevant features to gender recognition are selected from the first four formant frequencies and twelve MFCCs and feed the SVM classifier. While the relevant features to age has been used with k-NN classifier for the age recognizer model, using MATLAB as a simulation tool. A special selection of robust features is used in this work to improve the results of the gender and age classifiers based on the frequency range that the feature represents. The gender and age classification algorithms are evaluated using 114 (clean and noisy) speech samples uttered in Kurdish language. The model of two classes (adult males and adult females) gender recognition, reached 96% recognition accuracy. While for three categories classification (adult males, adult females, and children), the model achieved 94% recognition accuracy. For the age recognition model, seven groups according to their ages are categorized. The model performance after selecting the relevant features to age achieved 75.3%. For further improvement a de-noising technique is used with the noisy speech signals, followed by selecting the proper features that are affected by the de-noising process and result in 81.44% recognition accuracy.**

***Index Terms*—Age classification from speech, gender classification from speech, MFCC based gender and age recognition, SVM classifier.**

I. Introduction

The problems to be faced in the process of automatic speech-based age estimation are: Firstly, a large balanced database for different speaker age ranges is needed. Secondly, speech contains other speaker characteristics, including the speaker's weight, height and emotional condition, these characteristics may interact with age. (Bahari and Van hamme, 2011; Dobry et al. 2011; Florian, et al., 2007).

ARO-The Scientific Journal of Koya University
Volume III, No 2(2015), Article ID: ARO.10072, 06 pages
DOI: 10.14500/aro.10072
Received 11 February 2015; Accepted 03 July 2015
Regular research paper: Published 06 October 2015
Corresponding author's e-mail: fatima.faek@su.edu.krd

Automatic gender and age classification from speech can be performed using different approaches. For instance cepstral features, like Mel frequency cepstral coefficients (MFCC), are used by Florian, et al. for age recognition (Florian, et al., 2007). MFCC is reported to produce poor results for gender and age classification with recorded signals, either by using different microphones or in different places (noisy, and non-noisy) (Tiwari, et al., 2011). To avoid this problem, Sas, et al enhanced the MFCC features by analyzing the parameters that affect the process of extracting the features. Additionally the impact of these parameters on the gender recognition accuracy is studied (Sas, et al., 2013). Other researchers have used pitch or prosodic features with MFCC together; this improves the results of the gender and age recognizer (Harnsberger, et al., 2008).

While (Golfer and Mikes, 2005) studied the automatic gender classification from the speech signals of adult speakers using features related to the fundamental frequency (F0) and the first four formant frequencies**.** This approach is very active for gender classification of three classes, adult males, adult females, and children without gender discrimination, since the F0 and the formant frequencies, (especially the second and third formants), of children are higher than that of adults, i.e. the range of formant decreases with age (Potamianos and Narayanan, 2003). Female range changes are more gradual than male and the main differences become more significant after age twelve (Potamianos and Narayanan, 2003).

Different classifiers can be used for gender and age classification; Gaussian mixture models (GMM), hidden Markov models (HMM), support vector regression (SVR), and multilayer perceptrons (MLP) are proposed and tested by Hugo, et al., for gender classification, where the result was 95%. This result is for male/female classification and does not concern children (Hugo and Isabel, 2011). Sedaaghi used different classifiers for age estimation, like support vector machine (SVM), k-nearest neighbor (k-NN), probabilistic neural network (PNN), and GMM, when a result of 88% was obtained (Sedaaghi, 2009). Mirhassani, et al., used single layer feed forward neural network to estimate the age of children, and they used fuzzy data fusion to make the overall decision (Mirhassani, et al., 2014). Thomas et al combines three MLP outputs trained with Perceptual Linear Prediction (PLP)

features (Thomas, et al., 2014).

As any pattern recognition task the model performance depends on the robustness of the features and the classifier.
Many features are suggested by previous studies. However, this work focuses on the use of MFCC and formant frequencies. The work investigates the contribution of different MFCC and formant frequencies in age and gender recognition, especially in terms of the contribution of different frequency bands represented in their relative formants and MFCCs. The goal of this work is to use a few number of MFCC and formant frequency features, which are relevant features for gender and age classification.

In this work, the support vector machine and the k-NN are used as classifiers. SVM is one of the popular on-the-shelf classification method for its ability to separate the classes by an optimized hyper plain since it maximize the margin distance from the separator hyper plain to the support vector.

The data used in this work is a mix of noisy and clean speech signals, therefore this work follows a pre-processing (de-noising) technique and feature selection among the 12 MFCC, that are affected by the de-noising process for further improving the result of the age classifier (Faek and Al-Talabani, 2013).

II. The Database

The database consists of recorded speech sentences uttered by 114 Kurdish speakers (males and females), their ages lying between five and sixty-five years. This data has been recorded and collected by the author. The age recognizing algorithm is evaluated using the 114 speech samples after distributing the database in the following manner:

a) Children: less than 13 years (32 speakers).
b) Young: 14-19 years, (15 male speakers), and (15 female speakers).
c) Adults: 20-55 years, (17 male speakers), and (19 female speakers).
d) Seniors: up to 56 years, (9 male speakers), and (7 female speakers).

The two classes gender recognizer is evaluated using 82 adult males (40 speakers), and females (42 speakers), whereas, the 114 speech samples are used for the evaluation of the three classes gender recognizer, adult males (40 speakers), adult females (42 speakers), and children (32 speakers).

III. The Gender Classifier

In this work, the robust features among the first four formant frequencies are used as features for gender classification. Formants are the peaks of the speech signal spectrum. Formant frequencies are an acoustic resonance of the human vocal tract which is measured as an amplitude peak in the frequency spectrum of the sound. Extraction of Formant frequencies is achieved using linear prediction coefficients (LPC) based formants estimation technique as shown in Fig. 1 (Golfer and Mikes, 2005).

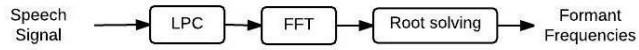

Fig. 1. Block diagram of extracting the formant frequency features.

The robust MFCC features are combined with the aforementioned features, since the MFCCs are the standard features in speech processing. They present the energy distribution of the speech signal in the frequency domain. This method is based on the Mel frequency scale and is related to human hearing. The technique of extracting MFCC features can be described by the block diagram shown in Fig. 2.

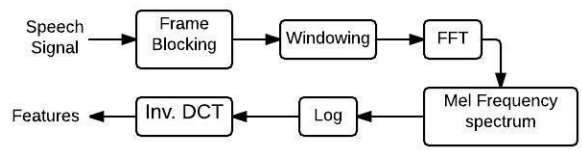

Fig. 2. Block diagram of extracting the MFCC features.

In the frame blocking section, the speech signal is divided into frames. Windowing minimizes the discontinuities present in the signal. The FFT is used for conversion of each frame from the time domain to the frequency domain which gives information about the energy rate at each frequency band. The Mel frequency spectrum is obtained by arranging the spectrum according to human hearing. Human hearing does not follow the linear scale but the Mel-spectrum scale which is a linear spacing below 1000 Hz and logarithmic scaling above 1000 Hz. Finally, the Mel-spectrum is converted back to the time domain by using the inverse DCT (Tiwari, et al., 2011).
Cepstral features are calculated from the log filter bank amplitudes (mj), as shown in Fig. 3. These features are calculated using the discrete cosine transformation as expressed below:

$$ci = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} mj \cos\left(\frac{\pi j}{N}(j-0.5)\right) \quad (1)$$

N is the number of filter bank channels which is set to 24. The SVM, with linear and non-linear separation function (LSF and NLSF) is tested as a classifier for two classes (adult males and females) gender recognition. SVM is a classifier that constructs an N-dimensional hyper-plane that separates the data optimally into two classes. With the SVM, the weights of the network are found by solving a quadratic programming problem; the separation function between classes in SVM may be linear, or non-linear. (Bocklet, et al., 2008; Santosh, et al., 2012).

http://dx.doi.org/10.14500/aro.10072

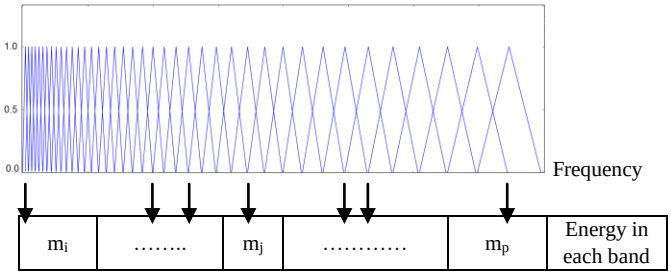

Fig. 3. Frequency response of a typical Mel-scaled triangular filter bank.

The goal of SVM modelling is to find the optimal hyper-plane that separates clusters of vector, (a set of features that describes one class), in such a way that the features that belong to the first class will be on one side of the plane and the features of the other class will be on the other side of the plane. The data close to the hyper-plane are the support vectors (Bocklet, et al., 2008; Santosh, et al., 2012).

IV. The Age Classifier

The robust formant frequencies in addition to the robust MFCCs to age classification are used as features. MFCCs are the best performing features for age recognition, but the drawback of these features is that they are not suitable for analyzing noisy signals (Tiwari, et al., 2011). So, for a further improvement of the age recognizer, a wavelet based, de-noising algorithm is followed to clean up the noisy speech signals as a pre-processing method (Faek and Al-Talabani, 2013), selecting robust features after the de-noising process is also of prime importance, to obtain the best results.

The k-Nearest-Neighbors (k-NN) is used as classifier, since it is a simple arbitrary classifier. This classifier is highly applicable in many cases. Simply this classifier classifies each set of the data in sample into one of the groups in training (Faek and Al-Talabani, 2013).

V. Feature Selection

In this work, the different features are analyzed, and the robust features are selected based on the frequency range that the feature represents. This process leads to obtain a good classification results with a few number of features. According to previous researchers, formants 2 and 3 differ from children to adults and from adult males to adult females (Potamianos and Narayanan, 2003), so these two formant features hold gender information and can be selected for better results, and the mid-frequency MFCC features are selected in this work for the same reason.

On the other hand, unlike the mid-frequency band; low and high frequency bands of a speech signal hold age information, especially the high frequency components of human voice, for they decrease with age. So based on this knowledge the low and high MFCCs are selected in this work, in addition to formants 1 and 4 for better age recognition results.

VI. Results and Discussion

A. Gender Classification Results

The results of the gender recognition of two classes(adult males and females), are shown in Table I, and Figs. 4, 5, 6, 7, 8 and 9, using different formant frequencies (F1, F2, F3 and F4) as features and SVM with linear separation function as a classifier. This algorithm is applied on a data consisting of 82 speakers (adult males and females).

TABLE I
Gender Recognizer Results of Tow Classes Using Different Formant Frequencies as Features, and SVM with Linear Suppuration Function as a Classifier

Feature	Recognition rate %
F2 and F4	85%
F3	89%
F2 and F3	90%
F3 and F4	89%
F1, F2 and F3	90%
F2, F3 and F4	89%
F1 and F3	90%
F1 and F4	81%
F1 and F2	83%
F1, F2, F3 and F4	88%

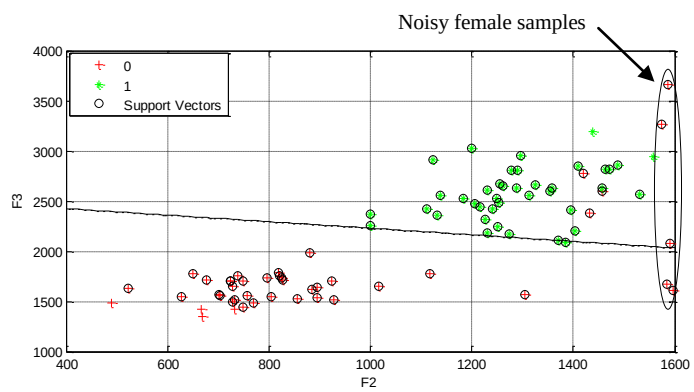

Fig. 4. Gender classification result using F2, F3 and SVM with LSF.

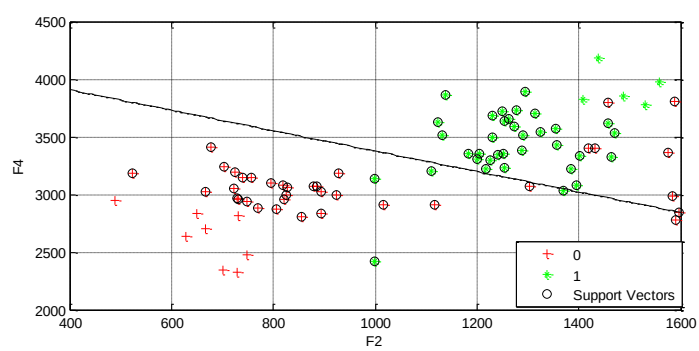

Fig. 5. Gender classification result using F2, F4 and SVM with LSF.

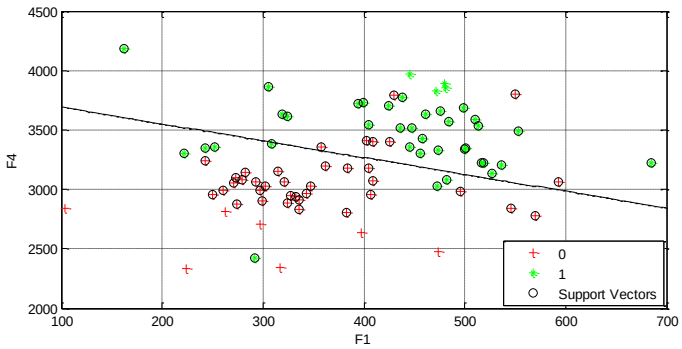

Fig. 6. Gender classification result using F1, F4 and SVM with LSF.

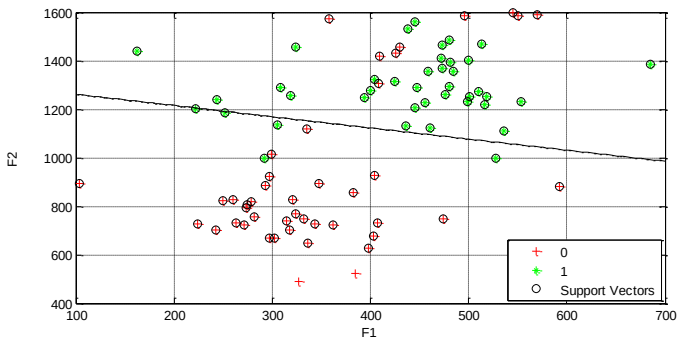

Fig. 7. Gender classification result using F1, F2 and SVM with LSF.

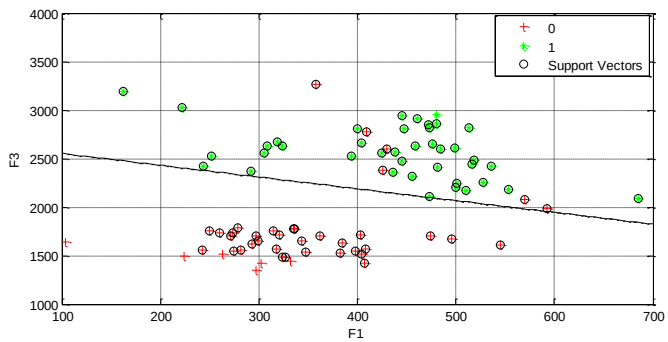

Fig. 8. Gender classification result using F1, F3 and SVM with LSF.

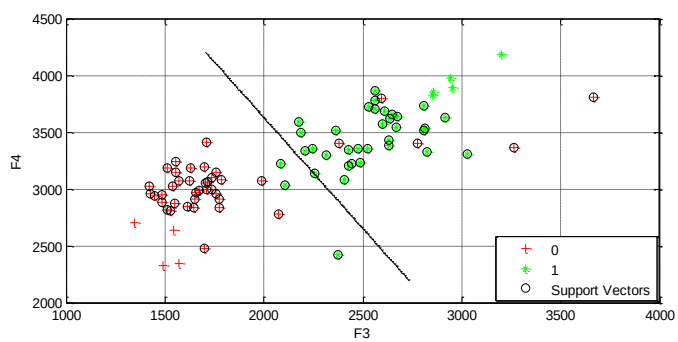

Fig. 9. Gender classification result using F3, F4 and SVM with LSF.

When using the SVM with non-linear separation function, the results of the gender recognition of two classes are as shown in Table II and Figs. 10, 11, 12, 13, 14, and 15.

TABLE II
GENDER RECOGNIZER RESULTS OF TOW CLASSES USING DIFFERENT FORMANT FREQUENCIES AS FEATURES, AND SVM WITH NON- LINEAR SUPPURATION FUNCTION AS A CLASSIFIER

Feature	Recognition rate %
F2 and F4	90%
F3	94%
F2 and F3	96%
F3 and F4	92%
F1, F2 and F3	91%
F2, F3 and F4	94%
F1 and F3	87 %
F1 and F4	83 %
F1 and F2	85 %
F1, F2, F3 and F4	90%

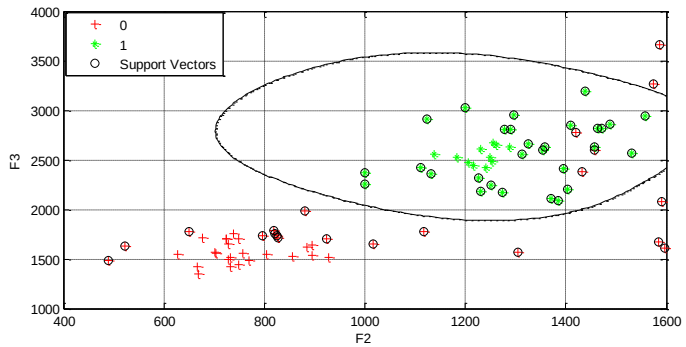

Fig. 10. Gender classification result using F2, F3 and SVM with NLSF.

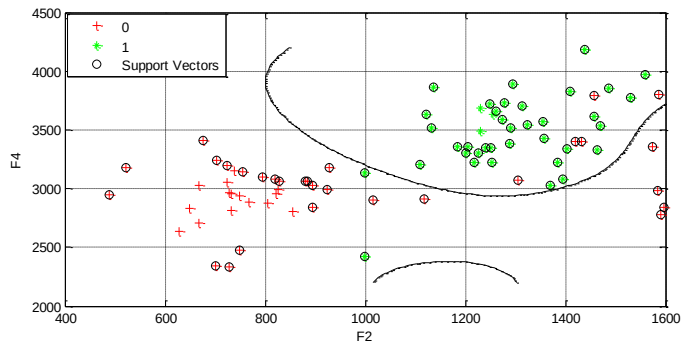

Fig. 11. Gender classification result using F2, F4 and SVM with NLSF

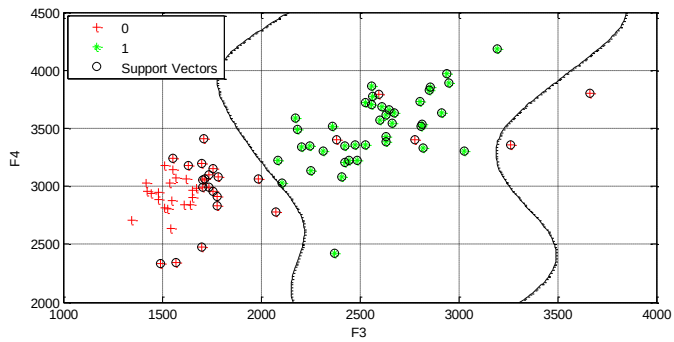

Fig. 12. Gender classification result using F3, F4 and SVM with NLSF.

The noise affects the formants of both genders but the formants of females change (increase) more than that of males. This problem appeared when using SVM with linear separated

http://dx.doi.org/10.14500/aro.10072

function; Non-linear separated function is more active with noisy signals, especially when using F2 and F3 as shown in Fig. 10.

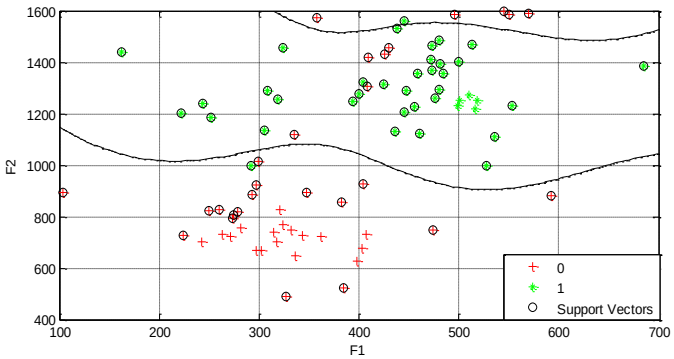

Fig. 13. Gender classification result using F1, F2 and SVM with non-linear separation function.

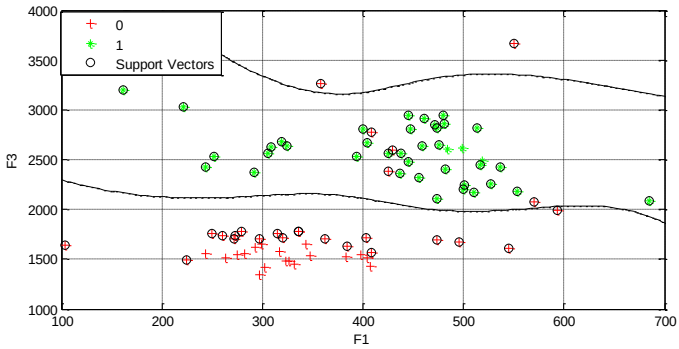

Fig. 14. Gender classification result using F1, F3 and SVM with NLSF.

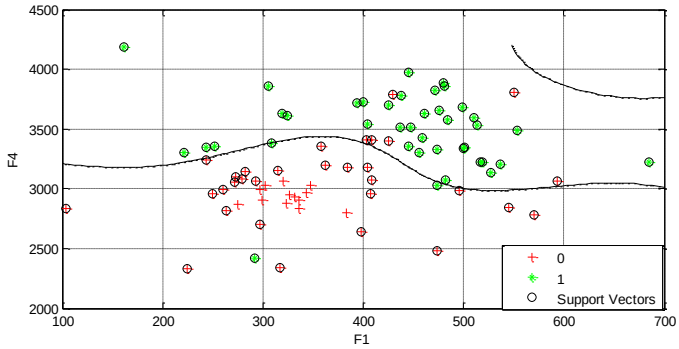

Fig. 15. Gender classification result using F1, F4 and SVM with NLSF.

It is clear that the best result is obtained using the second and third formants as feature.

Combining the MFCC with the formant frequencies as features, the results shown in Table III are achieved. The SVM with non- linear separation function is used as a classifier.

MFCCs (4-10) represent the mid frequency MFCC features. So it can be expected that these features in addition to formant 2 and 3 hold gender information, since they give the best results.

The results of the three classes' gender recognition are shown in Table IV, using different formant frequencies and different MFCC features, as well as the k-NN classifier. This algorithm is applied to (114) speakers (children without gender discrimination, adult males, and females).

Again the best result is obtained using F2, F3 and the mid MFCC features.

TABLE III
GENDER RECOGNIZER RESULTS OF TOW CLASSES USING MFCC AND FORMANT FREQUENCY AS FEATURES, AND SVM WITH NON- LINEAR SUPPURATION FUNCTION AS A CLASSIFIER

Feature	Recognition rate %
12 MFCC + F1, F2, F3 and F4	88%
4-9 MFCC + F1, F2, F3 and F4	92%
4-9 MFCC	96%
4-10 MFCC	93%
3-9 MFCC	92%
12 MFCC + F2 and F3	88%
4-9 MFCC + F2 and F3	96%

TABLE IV
GENDER RECOGNIZER RESULTS OF THREE CLASSES USING MFCC AND FORMANT FREQUENCY AS FEATURES, AND K-NN AS A CLASSIFIER

Feature	Recognition rate %
12 MFCC + F1, F2, F3 and F4	85%
F2 and F3	94%
12 MFCC + F2 and F3	85%
F1, F2 and F3	93%
F1 and F4	79%
4-9 MFCC + F2 and F3	94%
4-9 MFCC	94%
F1, F2, F3 and F4	92%

B. *Age Classification Results*

The results of the age classifier, using the first formant frequencies, and the 12 MFCC as features, and k-NN as a classifier, are shown in Table V. This algorithm is applied on a database of 114 speakers of different ages; five to sixty-five years. It is clear from the results of this part that F1 and F4 are important features for age classification, especially when combing them with the low and high frequency MFCC features. As a result, it can be expected that the low and high frequency features, (formants and MFCC), carry age information other than mid-frequency features.

TABLE V
AGE RECOGNIZER RESULTS USING MFCC AND FORMANT FREQUENCY AS FEATURES, AND K-NN AS A CLASSIFIER

Feature	Recognition rate %
F1and F4	66%
F2 and F3	52%
F2, F3 and F4	52%
F3 and F4	59%
F1, F2 and F3	54%
F1, F3 and F4	68%
F1, F2, F3, F4 and12 MFCC	69%
F1, F4 and 12 MFCC	57%
F1, F4 and MFCC 1, 2, 3, 4, 8, 9, 10, 11 and 12	75.3%

It can be observed from the results of gender recognition shown in Tables III, and the above results that MFCC's 4, 8 and 9 are useful for both gender and age classification.

The results of the age classifier, after de-noising the noisy speech signals, and using the aforementioned features and classifier are shown in Table VI. In this case, F1 and F4 are also of prime importance and the high frequency MFCC features have a great effect on the results, since they are affected by the de-noising technique used in this part.

The time cost for the all experiments done in this paper using SVM, is not exceeding 7 seconds. While in the case of k-NN the time cost is not more than 1.5 second. In both cases the complexity is not assigned as a serious drawback.

TABLE VI
AGE RECOGNIZER RESULTS AFTER THE DE-NOISING PROCESS USING MFCC AND FORMANT FREQUENCY AS FEATURES, AND K-NN AS A CLASSIFIER

Feature	Recognition rate %
F1 and F4	67%
F2 and F3	56%
F2, F3 and F4	56%
F3 and F4	59%
F1, F2 and F3	62%
F1, F3 and F4	69%
F1, F2, F3, F4 and12 MFCC	76%
F1, F4 and 12 MFCC	78%
F1, F4 and MFCC 7, 8, 9, 10, 11 and 12	81.44%

VII. CONCLUSION

The best gender classification results are obtained in the case of two classes and three classes with the second and third formant frequencies as features, since these formants differ from children to adults and from adult females to males. The mid-frequency MFCC features also hold gender information, so these MFCC features can be selected for better gender classification results.

Since the database is a mix of clean and noisy recorded speech sentences, SVM with a non-linear separated function is more active with the noisy speech signals than the linear separated function.

The experiments done in this work show that F1 and F4 are more relevant to human age recognition than F2 and F3, consequently they were selected for the age recognition model.

The 12 MFCC are also important features for age recognition, especially the low and high MFCCs, as they hold age information. However, the drawback of these features is that they are affected by noise. De-noising criteria as a pre-processing method leads to better results especially after selecting the robust features that are affected by the de-noising process among the twelve MFCCs, which are the high frequency features, in addition to the F1 and F4 features.

The selected features in this work are picked according to a pre-knowledge scheme, and not selected automatically. Consequently the obtained results need to be generalized to other datasets.

REFERENCES

Bahari, M.H. and Van hamme, H., 2011. Speaker age estimation and gender detection based on supervised non-negative matrix factorization, In: IEEE, *IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications*, Italy, 28 September 2011. USA.

Bocklet, T., Maier, A. and North, E., 2008. Age Detection of Children in Preschool and primary School Age with GMM-Based Super vector and Support Vector Machines/regression. In: *11th International Conference, TSD 2008*, Bron, Czech Republic, 8-12 September 2008.

Dobry, G., Hetch, M., Avegal, M., and Zigel, Y., 2011. Super vector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal, *IEEE Trans. Audio, Speech and Language Processing*, 19(7), pp.1975–1985.

Faek, F.K., Al-Talabani, A.K., 2013. Speaker recognition from noisy spoken sentences, *International Journal of Computer Applications*. 70(20), pp.11-14.

Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J.G. and Littel, B., 2007. Comparison of four approaches to age and gender recognition for telephone applications, In: IEEE, *IEEE international conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, 15-20 April 2007. USA.

Golfer, M. and Mikes, V. 2005. The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels, *Journal of Voice*, 19 (4), pp.544-554.

Harnsberger, J.D., Shrivastav, R., Brown, W.S., Rothman, H. and Hollien, H., 2008. Speaking rate and fundamental frequency as speech cues to perceived age, *Journal of Voice*, 22(1), pp.58-69.

Hugo, M. and Isabel, T., 2011. Age and gender detection in the I-DASH project ACM, *Transactions on Speech and Language Processing*, 7(4), 16 pages. DOI 10.1145/1998384.1998387.

Li, M., Han, K.J. and Narayanan, S., 2012. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech and Language*, 27, pp.151-167.

Mirhassani, S.M., Zourmand, A. and Ting, H.N., 2014. Age Estimation Based on Children's Voice: A Fuzzy-Based Decision Fusion Strategy. *Scientific World Journal*, [online] Available at: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4070543/> [Accessed 22 November 2014].

Potamianos, A. and Narayanan, S., 2003. Robust recognition of children's speech. *IEEE Trans. Speech Audio Processing*, 11(6), pp.603–616.

Santosh, G., Bharti, G. and Mehrotra, S.C., 2012. Gender identification using SVM with Combination of MFCC, *Advances in Computational Research*, 4(1), pp.69-73.

SAS. J. and SAS., A., 2013. Gender recognition using neural network and ASR techniques, *Journal of medical information and technologies*, 22, pp.179-187.

Sedaaghi, M.H., 2009. A comparative study of gender and age classification in speech signals, *Iranian Journal of Electrical & Electronic Engineering*, 5(1), pp.1-12.

Thomas, P., Vahid, H., Isabel, T., Annika, H., Miguel, S., 2014, Speaker age estimation for elderly speech recognition in European Portuguese. In: *The 15th Annual Conference of the International Speech Communication Association - INTERSPEECH 2014*, Singapore, 14-18 September 2014.

Tiwari, V., Ganga, G., Singhai, J. and Azad, M., 2011. Wavelet based noise robust features for speaker recognition, *Signal Processing: An International Journal (SPIJ)*, 5(2), pp.52-64.