Dempster-Shafer Fusion based Gender Recognition for Speech Analysis Applications

Jamil Ahmad, Khan Muhammad, Soon il Kwon, Sung Wook Baik*
College of Electronics and Information Engineering
Sejong University, Seoul, Republic of Korea
*Corresponding author email: sbaik@sejong.ac.kr

Seungmin Rho
Department of Multimedia
Sungkyul University, Anyang, Republic of Korea
smrho@sungkyul.ac.kr

***Abstract*— Speech signals carry valuable information about the speaker including age, gender, and emotional state. Gender information can act as a vital preprocessing ingredient for enhancing speech analysis applications like adaptive human-machine interfaces, multi-modal security applications, and sophisticated intent and context analysis based forensic systems. In uncontrolled environments like telephone speech applications, the gender recognition system should be adaptive, accurate, and robust to noisy environments. This paper presents a reasoning method governed by Dempster-Shafer theory of evidence for automatic gender recognition from telephone speech. The proposed method uses mel-frequency cepstral coefficients with a support vector machine to generate the initial prediction results for individual speech segments. The reasoning scheme collects and validates results from support vector machine and treats convincing predictions as valid evidence. It is argued that the consideration of valid evidence in the reasoning process improves recognition performance by avoiding unconvincing classification results. Experiments conducted on large speech datasets reveal the superiority of the proposed gender recognition scheme for speech analysis applications.**

Keywords— gender recognition; speech forensics applications; Dempster-Shafer theory; evidence fusion

I. INTRODUCTION

Social interactions often exploit verbal and non-verbal cues during speech communications. Speech carry linguistic (words, phrases, etc.) and para-linguistic information like age, gender, emotions of the speaker. Such speaker information extracted from their speech can be used in a variety of applications ranging from simple speech or speaker recognition, to sophisticated speech analysis applications involving high level reasoning. The use of such information in a preprocessing stage has proven useful in many speech analysis applications [1]. For instance, gender dependent speech recognition models has proven to be more accurate than gender independent models. Similarly, gender information can effectively bisect the search space for a speaker recognition system by eliminating all the less probable speakers not belonging to the identified gender class, thereby improving efficiency. These applications demand efficient and accurate gender information from speech signals.

Accurate gender information extracted from telephone speech signals can be utilized in a number of applications [2]. For instance gender dependent speech coding, gender-dependent speech modeling, gender-adaptive human computer interfaces, and as a preprocessing stage for a sophisticated acoustic analysis for lie detection applications. Gender dependent emotions recognition has also been investigated with improved results [3].

In [4], Harb and Chen introduced a general audio classifier system for gender recognition. They used MFCC feature vectors with multiple trained neural network classifiers. A simple majority voting scheme was used to combine the outcomes of multiple classifiers. Harb et al. presented an improved audio classifier method [5] utilizing multiple features with multiple classifiers to perform gender recognition task achieving an accuracy of 90.0 %. Very limited data was used to test the performance of the system for telephone speech. A weighted summation fusion scheme was employed by Li et al. [6] to determine gender using acoustic and prosodic features. Accuracy of 91.9% was achieved by their system. It has been shown in numerous cases that the performance is considerably improved as a result of early and late fusion strategies [7]. The use of a dynamic fusion strategy in such situations can considerably improve performance.

This paper presents a Dempster-Shafer theory based scheme for valid evidence fusion to improve gender recognition from telephone speech. The proposed strategy works more intuitively by giving more weights to high confidence predictions, and assigns less weight to low confidence predictions or simply ignores them. Such a dynamic weighting for the intermediate predictions makes it a more effective fusion scheme than other similar approaches. Details of the proposed reasoning scheme are provided in the subsequent sections.

II. METHODOLOGY

Typical gender recognition systems extract frequency features from speech signals. These features are used to build classification models for gender recognition. Statistical classifiers utilize the model to recognize speaker gender. In case of telephonic speech, the low quality of speech signals, unconstrained environment, background noise, and huge diversity of users make gender recognition a very challenging task. Robust systems for such situations involve extraction of robust features and carefully trained recognition models. An intuitive reasoning process can be utilized to improve the

*Corresponding author; Tel.: +82-02-3408-3797; Fax: +82-02-3408-4339

978-1-4673-8685-2/16/$31.00 ©2016 IEEE

recognition even further. A schematic diagram of the proposed reasoning framework for gender recognition is given in fig. 1.

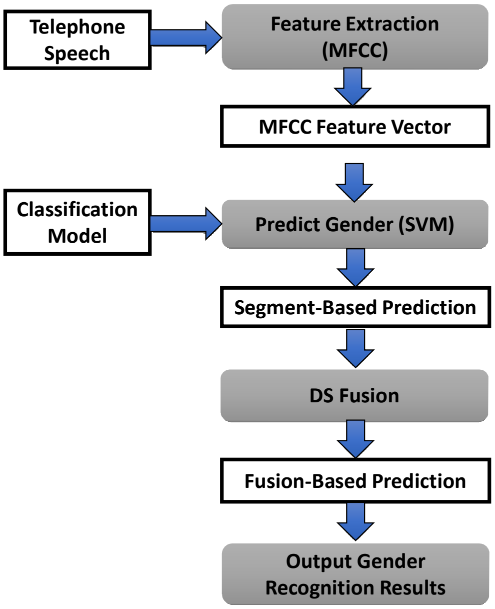


Fig. 1. Proposed reasoning framework for gender recognition

The entire framework is explained in detail in the following sub-sections.

A. *Features Extraction*

The most important phase in any computer-based recognition system is feature extraction [8]. They form the very basis upon which the whole recognition system depends. For gender recognition from speech, mel-frequency cepstral coefficients (MFCC) features have been widely used due to their robustness to noise [9, 10]. The incoming speech stream is segmented into small frames. The frames are categorized as speech and non-speech. Sixteen MFCC coefficients are extracted from each speech frame which makes the feature vector to be used for recognition.

B. *Gender Prediction*

Statistical classification schemes utilize features to determine decision boundaries between the distinct classes. These boundaries form the basis of recognition [11]. Several methods exist for the purpose of determining decision boundaries between classes. In our system, we utilized the MFCC features to train a support vector machine (SVM) classifier [12] with radial basis function kernel [13]. Instead of the class label, we used the probabilistic output of the SVM classifier in the reasoning process [14]. Since the frame is not sufficiently large to allow accurate recognition, several frame predictions were aggregated to compute a more stable segment based prediction result for both gender classes for use in the reasoning phase.

C. *Validity of Evidence*

Segment based gender classification results generated by SVM are considered as candidate evidence. However, instead of considering every prediction result as evidence, we impose a condition on the validity of the evidence based on the confidence level. Prediction results generated in the range 0.45 to 0.55 were considered as invalid evidence and ignored. It is argued that the low confidence results can mislead the overall recognition process towards misclassification. Therefore, ignoring unconvincing evidence can lead to better recognition performance. The validity condition for segment based predictions is given as:

$$(PS_M > 0.55 \wedge PS_F < 0.45) \vee (PS_F > 0.55 \wedge PS_M < 0.45) \qquad (1)$$

where PS_M and PS_F are the segment based prediction results for male and female subjects, respectively.

D. *Dempster-Shafer Fusion of Valid Evidence*

After performing evidence validity, the Dempster-Shafer fusion process is invoked to update belief values for the two gender classes. The first set of belief values are obtained from the first valid evidence observed. Subsequent valid evidence is combined with the previous belief values using the Dempster-Shafer fusion scheme. Valid positive evidence increase the belief value of the true class, whereas, the belief value for the true class gets reduced when negative evidence is encountered. The proposed fusion scheme is given as:

$$(PF_M^i, PF_F^i) = \left(\frac{PS_M^i.PF_M^{i-1}}{1-(PS_M^i.PF_F^{i-1} + PS_F^i.PF_M^{i-1})}, \frac{PS_F^i.PF_F^{i-1}}{1-(PS_M^i.PF_F^{i-1} + PS_F^i.PF_M^{i-1})} \right) \qquad (2)$$

Where PF_M and PF_F are the fusion based probabilities (or belief values) for male and female subjects, respectively. The updating of belief values is subject to the satisfaction of (1). An illustration of the evidence categorization during the course of a telephone call is given in fig. 2. Prediction results falling in the range [0.45 – 0.55] are ignored which improves recognition performance as well as causes the computations in the reasoning process to be reduced.

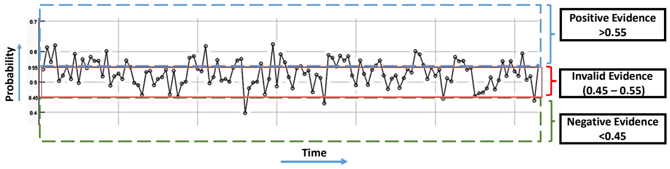


Fig. 2. Categorization as valid and invalid evidence

III. Experiments and Results

The experiments were conducted using a large speech dataset consisting of full recordings of Korean male and female subjects reporting emergency situations through telephone calls. The dataset contained about 65% male and 35% female voices. With the proposed scheme, we were able to achieve high confidence outputs for gender classes and miss-classifications were also reduced.

In order to assess performance of the reasoning based recognition scheme, confidence values are reported along with classification accuracies for some of the test subjects. Confidence values are computed by taking the mean of the intermediate belief values for all speech segments. Classification accuracy is computed by dividing the number of correctly classified subjects by the total number of test subjects.

$$ConfidenceValue = \frac{1}{N}\sum_{i=1}^{N} P \qquad (3)$$

where P represent the segment based prediction values for a speech segment and N are the number of speech segments.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (4)$$

Comparison between the confidence values for SVM only recognition scheme and the proposed scheme is given in fig. 3. It can be seen that the confidence values for the true class for all the test subjects are much higher than the SVM.

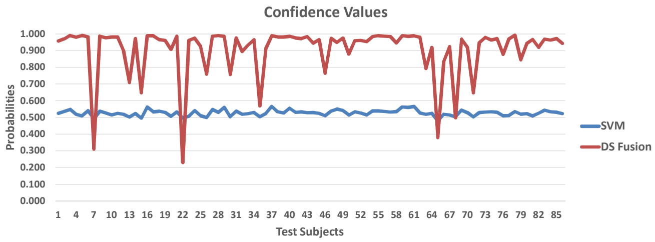
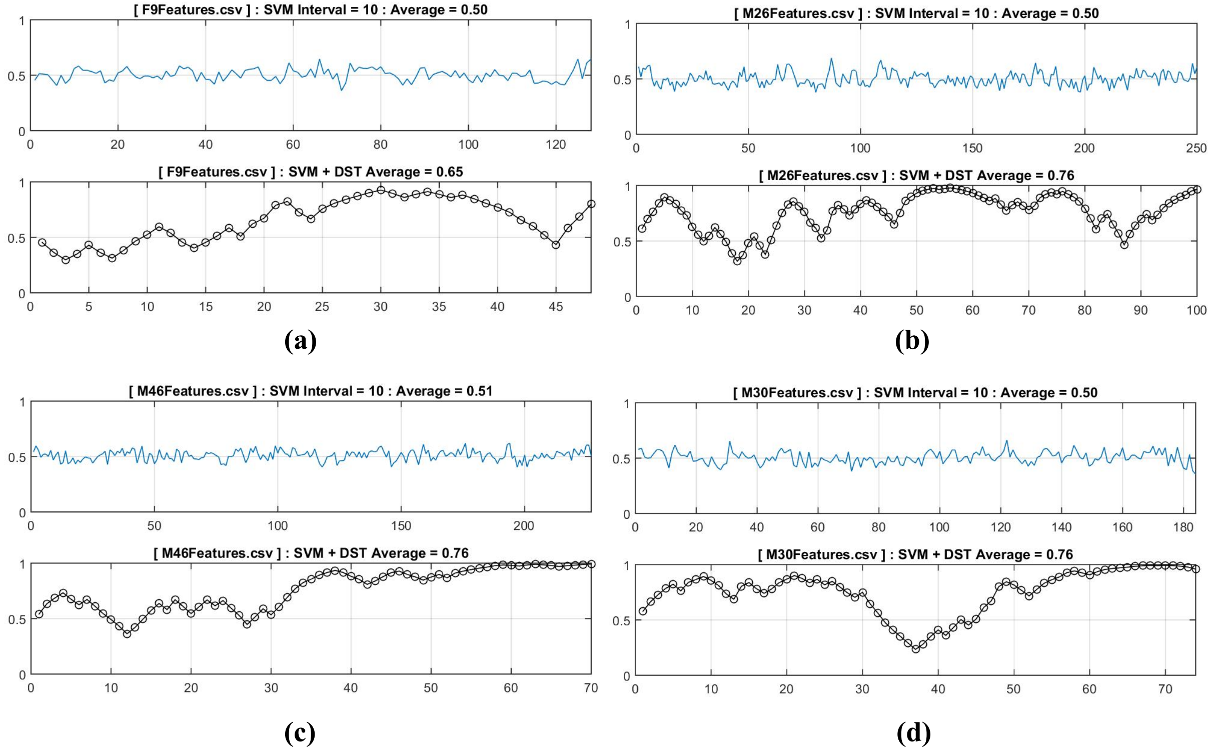


Fig. 3. Confidence values for SVM and SVM + DS Fusion

Fig. 3 reports the confidence values for some of the test subjects. The SVM predictions for almost all of the subjects are usually in the range [0.45 – 0.56]. SVM miss-classified six subjects, whereas, the proposed scheme was able to correctly classify two of those subjects correctly despite the fact that they were miss-classified by SVM. This improvement is achieved as a result of ignoring the unconvincing and probably miss-leading evidence. In order to get a deeper look into the predictions of fusion scheme, some of the results are shown in fig. 4.

It can be seen in figures 4a, 4b, and 4d, that the overall confidence value for SVM classification are 0.5. Which means that the SVM classifier is not able to discriminate among the two gender classes. In contrast to this, the proposed DS Fusion based scheme, generated correct prediction result for the true class with high confidence. The reason behind the improved performance is due to the consideration of valid evidence in the reason process and ignoring the unconvincing and misleading evidence.



(a)

(b)

(c)

(d)

Fig. 4. Comparison of predictions using SVM and SVM + DS Fusion (hard cases)

In fig. 4c, the SVM produces a correct prediction result with confidence 0.51. The confidence was elevated as and when positive valid evidence was encountered. All the sub-figures in fig. 4 represent prediction plots of the true class only.

The system's performance was evaluated using 2086 full length telephone calls recording from unconstrained real world environment. The proposed system correctly classified 1952 speakers out of 2086 yielding an overall accuracy of 93.6%. However, most of the miss-classified test samples contained severe background noise or were too short. For majority of the test samples, the system was able to identify gender correctly after accumulating 4 to 8 evidences. This early prediction of the proposed scheme makes it more useful to a wide variety of applications where such information could be utilized at a preprocessing stage.

Gender recognition performance primarily depends on speech stream lengths, and type of voice. Too short speech streams lack the sufficient amount of data required for making correct classification. Some voices are tough to identify because of their similarity to the opposite class. Emotional state of the speaker also affects the performance of such systems. Emotional states like happiness and anger greatly depends on the gender of the speaker. It becomes difficult to identify gender correctly when strong emotions are present in the speech [15].

A comparison of the proposed gender recognition system with other similar approaches is provided in table 1. It can be seen that the proposed system performs well for telephone speech compared to other schemes.

TABLE 1. Comparison with other gender recognition methods

Method	**Classification Accuracy (%)**
Pitch Features + NN [5]	90.0
General Audio Classifier [4]	91.7
Acoustic + Prosodic Fusion [6]	91.9
MFCC + SVM + DS Fusion (Proposed)	93.6

IV. Conclusions and Futute Work

In this paper, we presented a gender recognition strategy which employs Dempster-Shafer theory based reasoning process after the classification phase. The probabilistic output of SVM is analyzed, validated, used as evidence in the reasoning process. During the validation phase, unconvincing evidence is ignored and only valid convincing evidence is used in the reasoning process to update belief values for the two gender classes. The effectiveness of our approach is evident from the improvements witnessed in the results compared to the other technique. This work can be improved further if evidence from multiple heterogeneous and compatible classification schemes is fused to update belief values in a more complex reasoning process. Such a heterogeneous multi-classifier system can improve performance even further is used wisely. There is potential in the DS Fusion based scheme which will be explored further.

Acknowledgment

This work was supported by the ICT R&D program of MSIP/IITP. (No. R0126-15-1119, Development of a solution for situation-awareness based on the analysis of speech and environmental sounds).

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, *et al.*, "Emotion recognition in human-computer interaction," *Signal Processing Magazine, IEEE,* vol. 18, pp. 32-80, 2001.

[2] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Nöth, "Age and gender recognition for telephone applications based on gmm supervectors and support vector machines," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 1605-1608.

[3] T. Vogt and E. André, "Improving automatic emotion recognition from speech via gender differentiation," in *Proc. Language Resources and Evaluation Conference (LREC 2006), Genoa*, 2006.

[4] H. Harb and L. Chen, "Gender identification using a general audio classifier," in *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, 2003, pp. II-733-6 vol. 2.

[5] H. Harb and L. Chen, "Voice-based gender identification in multimedia applications," *Journal of intelligent information systems,* vol. 24, pp. 179-198, 2005.

[6] M. Li, K. J. Han, and S. Narayanan, "Automatic speaker age and gender recognition using acoustic and prosodic level information fusion," *Computer Speech & Language,* vol. 27, pp. 151-167, 2013.

[7] T. Bocklet, G. Stemmer, V. Zeissler, and E. Nöth, "Age and gender recognition based on multiple systems-early vs. late fusion," in *INTERSPEECH*, 2010, pp. 2830-2833.

[8] J. Ahmad, Z. Jan, and S. M. Khan, "A Fusion of Labeled-Grid Shape Descriptors with Weighted Ranking Algorithm for Shapes Recognition," *World Applied Sciences Journal,* vol. 31, 2014.

[9] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech communication,* vol. 55, pp. 237-251, 2013.

[10] T. Kinnunen, R. Saeidi, F. Sedlák, K. A. Lee, J. Sandberg, M. Hansson-Sandsten*, et al.*, "Low-variance multitaper MFCC features: a case study in robust speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 20, pp. 1990-2001, 2012.

[11] J. Ahmad;, M. Fiaz;, S.-i. Kwon;, M. Sodanil;, B. Vo;, and S. W. Baik, "Gender Identification using MFCC for Telephone Applications - A Comparative Study," *International Journal of Computer Science and Electronics Engineering,* vol. 3, pp. 351-355, 2015.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, pp. 273-297, 1995.

[13] S.-i. Amari and S. Wu, "Improving support vector machine classifiers by modifying kernel functions," *Neural Networks,* vol. 12, pp. 783-789, 1999.

[14] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST),* vol. 2, p. 27, 2011.

[15] D. Ververidis and C. Kotropoulos, "Automatic speech classification to five emotional states based on gender information," in *Signal Processing Conference, 2004 12th European*, 2004, pp. 341-344.