Fuzzy Based Multi-Source Data Fusion for Children's Age Estimation

Seyed Mostafa Mirhassani, Alireza Zourmand and Hua-Nong Ting

Department of Biomedical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia

Abstract— **Estimation of speaker's age is a challenge in speech processing area. This paper a novel approach for estimating a speaker's age is addressed. The method employs a "divide and conquer" strategy wherein the processing speech data are divided into six groups based on the vowel classes. Afterward, Mel-frequency cepstral coefficients are computed for each group and single layer feed-forward neural networks are applied to the features to make a primary decision. The extreme learning machine (ELM) method is used to train the classifiers. Subsequently, fuzzy data fusion is employed to provide an overall decision by aggregating the classifier's outputs. The results are then compared with vowel independent age estimation based on ELM and other well-known classification methods, including support vector machine and K-nearest neighbor. The processing speech data include six Malay vowels collected from 360 Malay children aged between 7 and 12 years. Experiments conducted based on six age groups revealed that fuzzy fusion of the classifier's outputs resulted in considerable improvement of up to 72.63% in age estimation accuracy. Moreover, the fuzzy fusion of decisions aggregated complimentary information of a speaker's age from varied speech sources.**

Keywords— **Fuzzy information fusion, extreme learning machine, age estimation, speech processing.**

I. Introduction

Speaker age has attracted considerable attention among researchers studying recent applications of speech processing. Speaker age provides valuable information that can also improve the performance of automatic speech recognition (ASR) systems as well.[1, 2] Many systems that employ speech data demand a type of user adaptation system that can be adapted with the age of a user. Additionally, in speech synthesis the appropriate language model can be properly selected based on the age information of the speaker. In commercial applications such as advertising, the target age group can be effectively selected based on speaker's age estimation. Moreover, in ASR systems, the underlying model can be adaptively selected to improve the speech recognition rate.

The estimation of a speaker's age is often performed based on groups of speakers in groups with a wider age range; however, few studies have conducted estimations based on children's speech. In this paper, the problem of age estimation in the context of children speech is addressed. In the diagnosis of some speech disorders, including dyslexia, the estimation of children's age provides valuable information [3, 4]. Moreover, in some interactive educational computer games [5,6], speech-based age estimation plays an important role in adapting systems to their users.

Based on different acoustical features and classifiers, a large number of methods for evaluation of speaker's age have been proposed in literature [2, 7]. Common features of such systems include using hidden Markov models (HMM), [8] support vector machines, [9] Gaussian mixture model (GMM), [2] and improvement of the age classes based on data projection to lower spaces [1]. Iseli et al.[1] modeled speakers by HMM weight supervector. Afterwards, to decrease the dimension of the input space, they employed a Weighted Supervised Non-Negative Matrix Factorization. Age of speakers has also been estimated based on Least Squares Support Vector Regression. Harnsberger et al. [10] investigated fundamental frequency and speaking rate to distinguish younger male speakers from older male speakers. Metz et. al. [9] used an SVM with RBF kernel, which received Mel-frequency cepstral coefficients (MFCC) and PLP coefficients as features. They repeated the experiments for different numbers of MFCCs. Muller and Burkhardt [7] proposed an age and gender estimation method based on a combination of regression and classification. They performed combination using the posterior probability of an SVM-based regressor trained depending on the speaker's age and a gender classifier.

Modeling of complicated distribution of training data in n-dimensional feature space require the use of higher order of nonlinearity or more complex modeling method. Such complexity results in problems that include overfitting of the classifiers. To cope with this problem, some approaches divided the complex problem into some simpler ones [11]. For this purpose, the processing data can be separated into subgroups so that a less complicated modeling method can efficiently handle the classification of each subgroup data. Through this approach, the fusion of decisions made by each preliminary classifier can be used to determine the overall classification results [12, 13, 14, 15].

For the purpose of age estimation based on speech data, we employ fuzzy data fusion in the current study in order to aggregate the decisions made by a few classifiers. A "divide and conquer" strategy is employed, in which the processing speech data are divided into some groups based on the vowel classes. There are two reasons behind this strategy.

First, decreasing the complicated distribution of the processing data improves the classifier's learning performance. Second, different vowel classes contain complimentary sets of information for age estimation. In the next step, the classifiers are applied on each group to make a primary decision. Subsequently, fuzzy data fusion is employed to provide an overall decision by aggregating the classifier's outputs. The rest of the paper is organized as follows: Section 2 presents the fuzzy fusion, Section 3 presents the experiments, and Section 4 concludes the paper.

II. Fuzzy Information Fusion

A. Problem definition

Let us suppose an n-class classification problem provided by m different classifiers. For a given speech sample x, the output of classifier *i* is the set of numerical values given by

$\{\mu_i^1(x), \mu_i^2(x), ..., \mu_i^n(x)\}$,

where $\mu_i^j(x) \in [0, 1]$ denotes membership degree of sample x to class *j* provided by classifier *i*. The higher this value, the more likely it is that the speech sample fits class j. Based on the classifier, $\mu_i^j(x)$ can be represented by probability, posterior probability at the output of a neural network, membership degree at the output of a fuzzy classifier, and so on. Consequently, the set $\pi_i(x) = \{\mu_i^j(x), j = 1, ..., n\}$ can be considered as a fuzzy set. In speech processing context, for each speech sample (feature), *m* fuzzy sets are provided. Therefore, the inputs for fusion procedure include $\{\pi_1(x), \pi_2(x), ..., \pi_i(x), ..., \pi_m(x)\}$.

B. Information fusion based on fuzzy aggregation

Combining different sources of information to improve the overall decision, also known as information fusion, is an effective way to cope with decision making under conflicting circumstances. After formulating the uncertain data, including decision of classifiers into the fuzzy sets, fuzzy aggregation is required to achieve an overall decision. In order to aggregate the fuzzy sets, numerous combination operators have been proposed in literature, in which each operator has its own properties that can be useful depending on the in-hand problem. The operators are categorized in three including Conjunctive combination, Disjunctive combination, and Compromise combination.

Based on a classification proposed by Bloch in 1996, these operators are recognized as contextual dependent (CD) operators [1]. There are different criteria to distinguish the context in our problem, including the information about possible conflicts between the sources and the reliability of each source. The operators have been introduced under the possibility theory, [21] but they are applicable in fuzzy set theory as well. Here, considering the context, the operators are adapted to deal with the fusion of the classifier's output [14].

C. Obtaining the classifier's decisions and confidence measurement

As previously mentioned, combining different sources of information to improve the overall decision is the idea behind the current study. Different vowels uttered by each speaker provide diverse sources of information, which are employed for estimation of speaker's age. Dealing with the age estimation problem, two different classification scenarios are studied including vowel-based age estimation and vowel-independent age estimation methods. The former method is employed for classifier fusion while the latter method is only used for comparison to the fusion method.

a) Combination operator and decision fusion

A large number of combination operators have been proposed in literature. The combination operator we used in this study is known as "fuzzy-or" operator. It is a compromise combination operator expressed as

$$\mu_f^j(x) = \gamma \max_{i=1}^{m} (\min(w_i \mu_i^j(x), \delta_i^j)) + (1-\gamma)\frac{1}{m}\sum_{i=1}^{m} w_i \mu_i^j(x)\delta_i^j \quad (1)$$

where $\mu_i^j(x)$ denotes the *j*th output of the *i*th classifier, which is normalized according to outputs of *i*th classifier; w_i is the local confidence coefficient associated with the classifier's output; δ_i^j is the global confidence coefficient, which is 1 for the classifier that results in the highest classification accuracy for a specific age and is zero for the rest of the classifiers; μ_f^j denotes the fusion result; and γ is the compensation degree. For $\gamma = 1$, the fuzzy-or operator behaves as max-operator, and the behavior of the operator for $\gamma = 0$ is similar to the arithmetic average of the fuzzy memberships. The confidence coefficient, w_i, represents the reliability of each classifier's output for a given test sample. Here, w_i can be obtained as follows:

$$w_i(x) = \exp(-0.5(\frac{\left|1 - \frac{S_{max1} - S_{max2}}{S_{max1} - S_{min}}\right|}{\sigma})^2) \quad (2)$$

where S_{max1}, S_{max2}, and S_{min} are the highest, second highest and lowest amounts in the output vector, respectively, which are produced by *i*th classifier, $[\mu_i^1, \mu_i^2, ..., \mu_i^n]$. In addition, σ is the standard deviation

of the Gaussian membership function. As Eq. 2 indicates, for a given test sample, the decision of a classifier is reliable if the highest output representing the classifier's decision is considerably higher than other outputs of the classifier. Consequently, w_i takes a higher value for reliable classifiers.

After performing fusion of the decisions provided by the classifiers based on Eq. 1, a vector representing the overall decision is obtained. The highest value in the vector presents the winner class assigned to the test sample. Experimental Results

III. Experimental Results

The single-frame feature extraction method was used to extract MFCC from the speech samples. Speech samples included 6 Malay vowels collected from 360 children aged between 7 and 12. The frame length for this method was 55 ms. For each speech sample, 120 MFCCs were computed, including 40 static, 40 delta, and 40 delta-delta coefficients.

Single hidden layer feed-forward neural-networks (SLFNs) based on Extreme earning machine (ELM) method was used as classification method in this study. In this method, input weights of the SLFN are randomly selected and the output weights are analytically computed (see [17] and [18] for more details about ELM). Comparative experiments with SVM in previous works, [18, 19] have revealed that this method can provide competitive performance to SVM. Experiments were accomplished based on a 3-fold cross validation method. Neural networks based on the ELM method and different activation functions as well as different numbers of hidden neurons were used for classification.

Table 1: A comparative result of vowel independent age estimation

Classification Method	Accuracy (%)	Specification
ANN (ELM)	24.77	100 hidden neurons,
SVM	24.21	Linear kernel
KNN	23.47	Euclidean distance, number of nearest neighbors = 20

A. Estimation of the speakers uttered different vowels

In this part, ANN method based on ELM training was applied to the speech database, which contained samples from of the entire set of phonemes. For the purpose of comparison, similar experiments were performed using the SVM and KNN methods which are summarized in Table 1.

B. Vowel-based age estimation

In this part of the experiment, which was performed before the classification, the database was divided to the vowel groups. Then ELM method was applied to each group in order to perform the age estimation. Meanwhile, different activation functions were used for the classifier. Table 2 presents the summary of the vowel-based age estimation results. As can be seen, the ELM method with Hardlim activation function provides higher vowel-based age estimation accuracy compared with other activation functions for the vowels /a/, /e/, /i/, and /u/.

C. Fusion of the classifier's decisions

After collecting the decisions of the classifiers from the previous part, an overall decision can be made by fusing the classifier's outputs. Here, σ in the confidence coefficient was 0.05 and the compensation degree was 0.6. Table 2 presents the fusion results. As can be seen, considerable improvement of age estimation is achieved by applying the fusion. The results show that different vowels reflect complimentary information regarding age estimation and the best accuracy was obtained by using sigmoid activation function. Dividing the speech data into vowel groups can decrease the complexity of data distribution in n-dimensional feature space. Therefore, classifiers can be more effectively trained on each group of the vowels. Meanwhile, the fuzzy formulation of the uncertainties of the classifier's output could help realize this objective.

D. Comparisons with other age estimation method

Based on the method proposed by Mahmoodi et. al. [9], a baseline system for age estimation has been provided to perform comparison with the proposed method. Best accuracy for the baseline system was obtained based on linear kernel (and Gamma = 2) for the SVM classifier at 30.56%. The proposed method outperformed the baseline method by decreasing the complexity of the processing data.

Table 2: Results of vowel based age estimation based on different activation functions and fusions

	Vowel groups						Fusion
AF	/a/	/e/	/ə/	/i/	/o/	/u/	
Hardlim	28.06	23.61	19.72	22.78	23.89	20.83	71.67
Sigmoid	20.00	21.11	21.94	20.56	23.61	18.61	72.63
Sin	22.50	16.94	17.22	17.78	15.56	17.22	67.22

IV. Conclusion

The fusion of several classifiers trained by different sources is used to estimate speaker's age. In order to reduce the complexity of the data distribution in n-dimensional feature space, the speech data is divided into six vowel groups. The vowel-based age classification is performed. SLFNs are also used for classification. Subsequently, fuzzy information fusion is used to provide decision fusion of the classifiers trained in the previous step. The decision fusion achieves a considerable improvement compared with the classification accuracy of each group or vowel independent classification. The fuzzy aggregation of complimentary information, which is collected from different classifiers, provides a rich source of data for age estimation analysis.

Acknowledgements

The authors would like to thank the University of Malaya for funding this study under UMRG grant (RP016A-13AET).

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. M. Iseli, Y.-L. Shue, and A. Alwan, (2006) Age-and gender-dependent analysis of voice source characteristics. IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2006 Proceedings. pp. I-I.
2. N. Minematsu, M. Sekiguchi, and K. Hirose (2002) Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2002, pp. I-137-I-140.
3. M. B. Denckla and R. G. Rudel (1976) Rapid automatized'naming (RAN): Dyslexia differentiated from other learning disabilities, Neuropsychologia, 14:471-479.
4. A. J. Fawcett and R. I. Nicolson, (1995) Persistent deficits in motor skill of children with dyslexia, Journal of Motor Behavior, 27:235-240.
5. J. Mostow, S. F. Roth, A. G. Hauptmann, and M. Kane (1994) A prototype reading coach that listens, PROCEEDINGS OF THE NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 1994, pp. 785-785.
6. J. Mostow, S. F. Roth, and A. G. Hauptmann, (1995) Demonstration of a reading coach that listens, Proceedings of the 8th annual ACM symposium on User interface and software technology, 1995, pp. 77-78.
7. C. A. Müller and F. Burkhardt, (2007) Combining short-term cepstral and long-term pitch features for automatic recognition of speaker age, INTERSPEECH, 2007, pp. 2277-2280.
8. M. H. Bahari, (2012) Speaker age estimation using Hidden Markov Model weight supervectors, in Information Science, Signal Processing and their Applications (ISSPA), 2012 11th International Conference on, 2012, pp. 517-521.
9. D. Mahmoodi, A. Soleimani, H. Marvi, F. Razzazi, M. Taghizadeh, and M. Mahmoodi, (2011) Age estimation based on speech features and support vector machine, Computer Science and Electronic Engineering Conference (CEEC), 2011 3rd, pp. 60-64.
10. J. D. Harnsberger, R. Shrivastav, W. Brown Jr, H. Rothman, and H. Hollien, (2008) Speaking rate and fundamental frequency as speech cues to perceived age, Journal of voice, 22:58-69.
11. L. Rokach, (2005) Ensemble methods for classifiers, Data Mining and Knowledge Discovery Handbook, ed: Springer, pp. 957-980.
12. J. A. Benediktsson and I. Kanellopoulos, (1999) Classification of multisource and hyperspectral data based on decision fusion, IEEE Transactions on Geoscience and Remote Sensing, 37: 1367-1377.
13. G. Lisini, F. Dell'Acqua, G. Trianni, and P. Gamba, (2005) Comparison and combination of multiband classifiers for Landsat urban land cover mapping, IEEE International Geoscience and Remote Sensing Symposium. IGARSS'05. Proceedings. 2005, pp. 2823-2826.
14. M. Fauvel, J. Chanussot, and J. A. Benediktsson, (2006) Decision fusion for the classification of urban remote sensing images, IEEE Transactions on Geoscience and Remote Sensing, 44: 2828-2838.
15. G. Amici, F. Dell'Acqua, P. Gamba, and G. Pulina, (2004) A comparison of fuzzy and neuro-fuzzy data fusion for flooded area mapping using SAR images, International journal of remote sensing, 25:4425-4430.
16. X. Wang, J. Zhang, and Y. Yan, (2011) Discrimination between pathological and normal voices using GMM-SVM approach, Journal of Voice, 25:38-43.
17. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, (2004) Extreme learning machine: a new learning scheme of feedforward neural networks, Neural Networks, 2004. Proceedings. 2004 IEEE International Joint Conference on, 2004, pp. 985-990.
18. R. Zhang, G.-B. Huang, N. Sundararajan, and P. Saratchandran, (2007) Multicategory classification using an extreme learning machine for microarray gene expression cancer diagnosis, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 4:485-495.
19. T. Helmy and Z. Rasheed, (2009) Multi-category bioinformatics dataset classification using extreme learning machine, in Evolutionary Computation, CEC'09. IEEE Congress on, 2009, pp. 3234-3240.
20. I. Bloch, (1996) Information combination operators for data fusion: A comparative review with classification, Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, vol. 26:52-67.
21. D. Dubois and H. Prade, (1992) Combination of fuzzy information in the framework of possibility theory, Data fusion in robotics and machine intelligence, pp. 481-505.

Author: Dr. Hua-Nong TING
Institute: University of Malaya
Street: Jalan Pantai Baharu
City: Kuala Lumpur
Country: Malaysia
Email: tinghn@um.edu.my