Age and Gender Recognition from Speech Patterns Based on Supervised Non-Negative Matrix Factorization

Mohamad Hasan Bahari, and Hugo Van hamme
Centre for Processing Speech and Images, Katholieke Universiteit Leuven, Leuven, Belgium
{MohamadHasan.Bahari|hugo.vanhamme}@esat.kuleuven.be

In many criminal cases, evidence might be in the form of recorded conversations, possibly over the telephone. Therefore, law enforcement agencies have been concerned about accurate methods to profile different characteristics of a speaker from recorded voice patterns, which facilitate to identify him/her or at least narrow down the number of suspects. This paper proposes a new gender and age group recognition approach based on Non-Negative Matrix Factorization (NMF) (Lee and Seung, 2001). First, an acoustic model is trained for all speakers in a training database including male and female speakers of different age. Then, Gaussian Mixture (GM) weights are extracted and concatenated to form a supervector for each speaker. Finally, Supervised NMF (SNMF) is applied to detect the gender and age group of unseen test speakers. Evaluation results on a corpus of read and spontaneous speech in Dutch confirms the effectiveness of proposed scheme.

Corpora

In this research, speech patterns of 555 speakers from the N-best evaluation corpus (Van Leeuwen et al., 2009) were used. The corpus contains live and broadcast commentaries, news, interviews, and reports broadcast in Belgium. Table 1 shows the number of speakers in different age-gender categories. Speakers of the database are divided into training and testing data sets.

Table 1. The number of speakers in different age-gender categories.

Category Name	Young Male	Young Female	Middle Male	Middle Female	Senior Male	Senior Female
Age	18-35	18-35	36-45	36-45	46-81	46-81
Number of Speakers	85	53	160	41	191	25

Proposed Method

The proposed age-gender recognizer proceeds in two steps:

A. Training Phase:

The general architecture of the proposed method in the training phase is illustrated in Figure 1.
The acoustic features consist of MEL spectra with mean normalization and vocal tract length normalization (Duchateau et al. 2006), augmented with their first and second order time derivatives. These features are then mapped to a 36 dimensional space by means of a discriminative linear transformation and decorrelated (Demuynck, 2001).
The system uses Hidden Markov Model, composed of a shared pool of 49740 Gaussians to model the observations in 3873 cross-word context-dependent tied triphone states. All acoustic units –context-dependent variants of one of the 46 phones, silence, garbage and speaker noise– have a 3-state left-to-right topology.
The speaker dependent weights for each speaker of the training data result from a re-estimation of the speaker independent weights based on a forced alignment (using the speaker independent model) of the training data for that speaker.

After obtaining a speaker dependent model for all speakers of the training data set, GM weights are extracted and concatenated to form a supervector for each speaker. Then, all supervectors along with their corresponding age-gender category are applied to train a SNMF. The training procedure of a SNMF is introduced in (Van hamme, 2008).

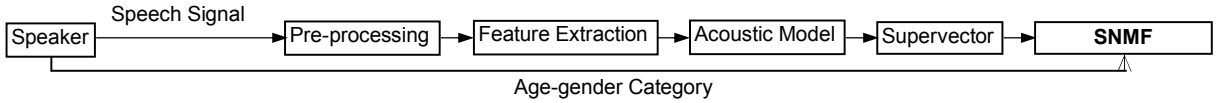

Figure 1 The architecture of proposed method in training phase.

B. Testing Phase

The architecture of the proposed method the in testing phase is shown in Figure 2.
During testing, the procedure of obtaining GM weights supervectors is repeated for each single speaker of the test data set. Then the resulting supervector is fed into the trained SNMF to determine its corresponding age-gender category.

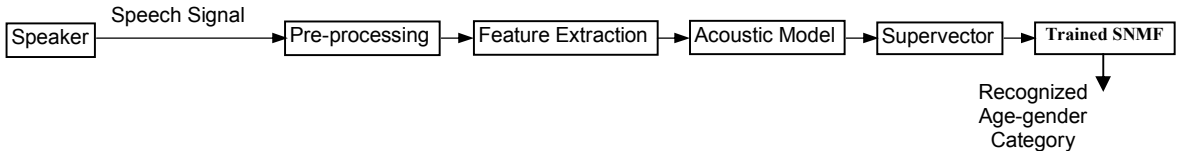

Figure 2 The architecture of proposed method in testing phase.

Results

A 5-fold cross-validation method was applied to test the proposed method over all 555 speakers of the database. The accuracy of the introduced method in gender recognition is 96%. Table 2 shows the average of age-gender group recognition accuracy over all performed experiments. The second row lists the prior class probability, or "chance levels". Hence, the proposed method performs better than guessing.

Table 2. Age group recognition accuracy in %.

Category Name	Young Male	Young Female	Middle Male	Middle Female	Senior Male	Senior Female
Prior	**15**	10	29	7	34	4
Accuracy	13	**77**	**44**	**24**	**76**	**16**

References

Demuynck, K. (2001). Extracting, Modelling and Combining Information in Speech Recognition, *Ph.D. thesis, Katholieke Universiteit Leuven*.

Duchateau, J., M Wigham, K Demuynck, and H Van hamme. (2006), A flexible recogniser architecture in a reading tutor for children. *ITRW on Speech Recognition and Intrinsic Variation*, 59–64.

Lee, D. D., and H. S. Seung. (2001). Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 556–562.

Van hamme, H. (2008). HAC-models: a novel approach to continuous speech recognition. *Interspeech*, 2554-2557.

Van Leeuwen, D. A., et al. (2009). Results of the n-best 2008 dutch speech recognition evaluation. *Interspeech*, 2571-2574.