Speaker age estimation and gender detection based on supervised Non-Negative Matrix Factorization

Conference Paper · October 2011

DOI: 10.1109/BIOMS.2011.6052385 · Source: IEEE Xplore

2 authors:

Mohamad Hasan Bahari
Sensifai
45 PUBLICATIONS **522** CITATIONS
SEE PROFILE

Hugo Van hamme
KU Leuven
274 PUBLICATIONS **3,366** CITATIONS
SEE PROFILE

Speaker Age Estimation and Gender Detection Based on Supervised Non-Negative Matrix Factorization

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

Abstract— **In many criminal cases, evidence might be in the form of telephone conversations or tape recordings. Therefore, law enforcement agencies have been concerned about accurate methods to profile different characteristics of a speaker from recorded voice patterns, which facilitate the identification of a criminal. This paper proposes a new approach for speaker gender detection and age estimation, based on a hybrid architecture of Weighted Supervised Non-Negative Matrix Factorization (WSNMF) and General Regression Neural Network (GRNN). Evaluation results on a corpus of read and spontaneous speech in Dutch confirms the effectiveness of the proposed scheme.**

Keywords-age estimation; gender detection; weighted supervised non-negative matrix factorization; general regression neural network

I. INTRODUCTION

Law enforcement agencies have been concerned about different biometric techniques to confirm the identity of an individual [1]. Different biometric characteristics can be used for forensic identification such as fingerprint patterns, face characteristics, hand geometry, signature dynamics and voice patterns [1]. Choosing a method depends on its reliability in a particular application and the available data.

In some criminal cases, available evidence might be in the form of recorded conversations. Speech patterns can include important information for law enforcement personnel [2]. For example, a person's speech pattern can provide information about his/her age, gender, dialect, emotional or psychological state and membership of a particular social or regional group. Therefore, speech can be used for speaker identification which is highly demanded in many cases such as kidnapping, threatening calls and false alarms [2].

In this research, we focus on speaker gender detection and age estimation. Since perceptions of gender and age have a significant mutual impact on each other, these two characteristics are studied together in many publications [3-4]. Computerized speech-based age estimation is difficult from different points of view. First, usually there exists a difference between the age of a speaker as perceived, namely the perceptual age, and their actual age, namely the chronological age. Second, developing a robust age recognition system requires a labeled, wide age-range and balanced database. Third, voice patterns are affected by many parameters, such as weight, height and emotional condition, i.e. there is a significant intra-speaker variability that is not related to or only correlated with age.

The problem of age group recognition has been addressed previouosly [3-6]. For example, Bocklet and his colleagues introduced a method based on a GMM mean supervector and a Support Vector Machine (SVM) to classify speakers into seven age-gender categories [3]. They used Mel Frequency Cepstral Coefficients (MFFCs) as features in their recognizer. Although this method was attractive from several aspects, it demands working with very large dimensions if the number of Gaussians in GMM be high. In [7], the GMM universal background model is merged with the SVM classifier and the problem of high dimensional supervectors is tackled by using Gaussian mixture weight supervectors, which have a lower dimension compared to mean or variance supervectors. Zhang et. al. reported age and gender recognition results with the use of an unsupervised Non-negative Matrix Factorization (NMF) over Gaussian mixture weight supervectors in [8]. In their approach, the acoustic features consist of Mel Spectra with mean normalization and Vocal Tract Length Normalization (VTLN) [9], augmented with their first and second order time derivatives. Although their method could recognize the gender of speakers with high accuracy, it is not very successful for age estimation. They also conclude that adding VTLN decreases the accuracy of gender detection but it helps in age recognition.

In this paper, we introduce a new gender detection and age estimation approach. To develop this method, after determining an acoustic model for all speakers of the database, Gaussian mixture weights are extracted and concatenated to form a supervector for each speaker. Then, a hybrid architecture of WSNMF and GRNN is developed using the acquired supervectors of the training data set. The obtained hybrid method is applied to detect the gender of test set speakers and to estimate their age.

This paper is organized as follows. Section 2 introduces WSNMF and GRNN. In section 3, the proposed approach is elaborated. The evaluation results are illustrated in section 4. The paper finishes with a conclusion in section 5.

This work is supported by the European Commission as a Marie-Curie ITN-project (FP7-PEOPLE-ITN-2008), namely Bayesian Biometrics for Forensics (BBfor2), under Grant Agreement number 238803.

II. BACKGROUND

In this section, the applied mathematical tools including WSNMF and GRNN are briefly introduced.

A. WSNMF

NMF is a popular machine learning algorithm [10], which is successfully applied a.o. to word recognition [11], sound source separation [12] and spam filtering [13]. During the last decade, different extensions of NMF such as Supervised NMF (SNMF) [14] and Weighted NMF (WNMF) [15] have been developed to solve real world problems. In this paper, the idea of WNMF is merged with SNMF and results in WSNMF.

1) SNMF

The problem addressed by SNMF is defined as follows. Assume that we are given a training data set $S^{tr1}= \{(x_1, y_1), \ldots, (x_n, y_n), \ldots, (x_N, y_N)\}$, where x_n denotes a vector of observed characteristics of the data item and y_n denotes a label vector, i.e. a vector containing one in the i-th row if x_n belongs to the i-th class and zeros eslewhere. A vector can be member of multiple classes, i.e. y_n can have multiple non-zero elements. The goal is to approximate a classifier function (g), such that for an unseen observation x^{tst}, $\hat{y}=g(x^{tst})$ is as close as possible to the true label.

If all elements of S^{tr1} are non-negative, this problem can be solved by SNMF directly. First, training data is used to form a general matrix V^{tr} as follows.

$$\begin{aligned} V_S^{tr} &= \begin{bmatrix} y_1 & \cdots & y_N \end{bmatrix} \\ V_B^{tr} &= \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \\ V^{tr} &= \begin{bmatrix} V_S^{tr} \\ V_B^{tr} \end{bmatrix} \end{aligned} \tag{1}$$

Then, the non-negative matrix V^{tr}, which is of size $M \times N$, is decomposed into two new non-negative matrices, namely W^{tr} and H^{tr} of size M×Z and Z×N respectively.

$$\begin{aligned} \begin{bmatrix} V_S^{tr} \\ V_B^{tr} \end{bmatrix} &\approx \begin{bmatrix} W_S^{tr} \\ W_B^{tr} \end{bmatrix} H^{tr} \\ W^{tr} &= \begin{bmatrix} W_S^{tr} \\ W_B^{tr} \end{bmatrix} \end{aligned} \tag{2}$$

This factorization can be performed by minimizing the following extended Kullbeck-Leibler divergence.

$$\begin{aligned} &D_{KL}\left(V^{tr} \middle\| W^{tr} H^{tr}\right) = \\ &\sum_{mn} V_{mn}^{tr} \log\left[\frac{V_{mn}^{tr}}{\left(W^{tr} H^{tr}\right)_{mn}}\right] + \left(W^{tr} H^{tr}\right)_{mn} - V_{mn}^{tr} \\ &+ \rho \sum_{zn} \left(H^{tr}\right)_{zn} \end{aligned} \tag{3}$$

The last term penalizes large entries in H^{tr}, so ρ controls the sparsity of H^{tr}. It can be shown that the above-mentioned function is non-increasing under the following multiplicative updating rules.

$$\begin{aligned} W^{tr} &\leftarrow \mathrm{P}_1 \circ \mathrm{P}_2 \\ \mathrm{P}_1 &= \frac{\left[W^{tr}\right]}{\left[1_{M\times N} (H^{tr})^T\right]} \\ \mathrm{P}_2 &= \frac{\left[V^{tr}\right]}{\left[W^{tr} H^{tr}\right]} (H^{tr})^T \\ H^{tr} &\leftarrow \mathrm{T}_1 \circ \mathrm{T}_2 \\ \mathrm{T}_1 &= \frac{\left[H^{tr}\right]}{\left[(W^{tr})^T 1_{M\times N} + \rho\right]} \\ \mathrm{T}_2 &= (W^{tr})^T \frac{\left[V^{tr}\right]}{\left[W^{tr} H^{tr}\right]} \end{aligned} \tag{4}$$

where $A \circ B$ and $\frac{[A]}{[B]}$ are the element-wise product and division of matrixes A and B respectively, $1_{M\times N}$ is a matrix of size $M \times N$ with all elements equal to 1 and the sign ^{T} is the transpose operator.

Calculation of W^{tr} by factorizing the V^{tr} is called training the SNMF. As can be seen in the following relation, W^{tr} which was obtained from the training phase, is used to determine the class of unseen patterns, x^{tst}:

$$\hat{y} = g(x^{tst}) = W_S^{tr} \arg\min_{H^{tst}} D_{KL}\left(x^{tst} \middle\| W_B^{tr} H^{tst}\right) \tag{5}$$

Notice that ŷ returns a fuzzy class membership that requires a decision criterion, such as thresholding or selecting the maximum entry.

2) WSNMF

To adjust the importance of different elements of V^{tr}, weighted NMF is introduced by Ho and his colleagues [15]. In this research, we adopt this idea to emphasize on V_s^{tr} in the factorization process. In this method, the updating rules are as follows.

$$\begin{aligned} W^{tr} &\leftarrow \mathrm{P}_1 \circ \mathrm{P}_2 \\ \mathrm{P}_1 &= \frac{\left[W^{tr}\right]}{\left[L (H^{tr})^T\right]} \\ \mathrm{P}_2 &= \frac{\left[L \circ V^{tr}\right]}{\left[W^{tr} H^{tr}\right]} (H^{tr})^T \\ H^{tr} &\leftarrow \mathrm{T}_1 \circ \mathrm{T}_2 \\ \mathrm{T}_1 &= \frac{\left[H^{tr}\right]}{\left[(W^{tr})^T L + \rho\right]} \\ \mathrm{T}_2 &= (W^{tr})^T \frac{\left[L \circ V^{tr}\right]}{\left[W^{tr} H^{tr}\right]} \end{aligned} \tag{6}$$

where L is a matrix with the same size of V^{tr}, which is determined as follows.

$$\begin{aligned} L_S &= \beta[1_{S\times N}] \\ L_B &= [1_{B\times N}] \\ L &= \begin{bmatrix} L_S \\ L_B \end{bmatrix} \end{aligned} \tag{7}$$

where $1_{S\times N}$ and $1_{B\times N}$ are two matrices with the same size of $V_s^{tr}$and V_B^{tr} respectively and all of their elements are equal to one. β is a factor determining importance of the supervision information. A reasonable value for this factor, which is also used in this paper, is

$$\beta = \frac{\sum_{bn} \left(V_B^{tr}\right)_{bn}}{\sum_{sn} \left(V_S^{tr}\right)_{sn}} \tag{8}$$

B. GRNN

A GRNN is used in conjunction with WSNMF to estimate the age of the speakers. A GRNN is a universal function approximator and was introduced in [16]. It is an approach with a one-pass learning algorithm such that for a given training data set S^{tr2}= {(a_1, b_1), . . ., (a_k, b_k), . . . , (a_K, b_K)}, the regression function is

$$\begin{aligned} F(a) &= \frac{\sum_{k=1}^{K} b_k \exp(-\frac{d(a-a_k)^2}{2\sigma^2})}{\sum_{k=1}^{K} \exp(-\frac{d(a-a_k)^2}{2\sigma^2})} \\ d(a-a_k)^2 &= (a-a_k)^T.(a-a_k) \end{aligned} \tag{9}$$

where σ is the standard deviation of the Gaussian kernel functions assumed around each sample. It is also called "smoothing parameter".

A GRNN has different advantages over other neural networks (NNs), which is the reason of applying it in this research:

- a GRNN does not require iterative learning algorithms. Instead, it has a one pass and fast learning. Standard supervised neural network architectures such as multilayer perceptrons and radial basis functions infer a parameterized model (the weights forming the parameters) from the available training data. These networks use the back-propagation algorithm for training, which may take a large number of iterations to converge, while global convergence cannot be assured.
- a GRNN requires only a fraction of the training samples that a back propagation based neural network would need. In other words, a GRNN can be effectively applied in the case of sparse data.

III. THE PROPOSED APPROACH

In this section, the proposed approach for gender recognition and age estimation is elaborated. To introduce this method, first the procedure of forming a supervector for a speaker is explained. Then, the proposed scheme in the training and testing phases is elucidated in details.

A. Feature selection, acoustic model and supervectors

The acoustic features consist of MEL spectra with mean normalization and vocal tract length normalization [17], augmented with their first and second order time derivatives. These features are then mapped to a 36 dimensional acoustic space by means of a discriminative linear transformation and are decorrelated [18]. The acoustic model uses a shared pool of 49740 Gaussians to model the observations in 3873 cross-word context-dependent tied triphone HMM states, each modeled with a Gaussian mixture (10). 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 mixture weights for each speaker of the training data set result from a re-estimation of the speaker independent weights based on a forced alignment of the training data for that speaker using a speaker-independent acoustic model. Subsequently, the Gaussian mixture weights are extracted and concatenated to form a supervector for each speaker.

For the n^{th} speaker consider the s^{th} Gaussian mixture with the following state probability density function (PDF).

$$f_s(o_t) = \sum_{j=1}^{J_s} w_j^s \Delta(o_t, \mu_j^s, \Sigma_j^s) \tag{10}$$

where O_t is the acoustic vector at time t, w_j^s is the mixture weight for the j^{th} component of the s^{th} mixture, Δ is a Gaussian probability density function with mean μ_j^s and covariance matrix $\sum_j^s$ and J_s is the total number of Gaussians in the s^{th} mixture. The supervector of Gaussian mixture weights can now be formed based on a Estimate-Maximize retraining of the mixture weights on data of speaker n only:

$$\begin{aligned} \lambda^s &= N^s \begin{bmatrix} w_1^s & \cdots & w_q^s & \cdots & w_Q^s \end{bmatrix}^T \\ \chi_n &= \begin{bmatrix} (\lambda^1)^T & \cdots & (\lambda^s)^T & \cdots & (\lambda^S)^T \end{bmatrix}^T \end{aligned} \tag{11}$$

where χ_n is the supervector of n^{th} speaker, S is the total number of mixtures, N^s is the number of frames observed in the s^{th} mixture during Viterbi alignment and Q is the total number of weights in the s^{th} mixture. Note that each element of λ^s is also equal to the sum over the training data of the posterior probabilities of the Gaussians or $(\lambda^S)_q = \sum_t \gamma(q,t)$.

B. Training phase

Speakers of the database are divided into two disjoint subsets namely training and testing data sets. The training

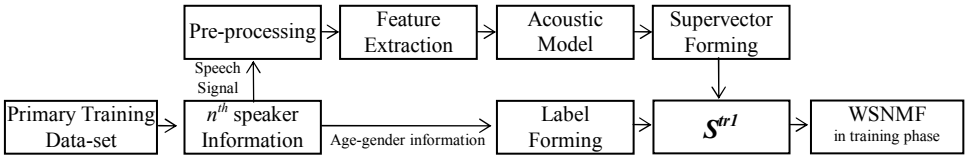

Figure 1. Block diagram of proposed method in primary training phase

patterns are also divided into primary training patterns, which are used for training the WSNMF, and secondary training patterns, which are used for training the GRNN.

1) Primary training

The general architecture of the proposed method in the primary training phase is illustrated in figure 1. As can be seen in this figure, first, the above-mentioned supervector forming procedure is applied to form a supervector for each speaker in the primary training data set. Then, an age-gender category label is formed for each supervector. Each label is a vector with dimension equal to the total number of considered age-gender categories (in this case six). The label of the n^{th} speaker, u_n, which belongs to the d^{th} category, is formed such that the d^{th} element of label vector is equal to 1 and the other elements are equal to zero. For example, if a speaker belongs to the second category, its label vector is $u=[0\ 1\ 0\ 0\ 0\ 0]^T$.

After calculating all N supervectors of the primary data set and labeling them by their age-gender category, we obtain $S^{tr1}= \{(\chi_1^{tr1}, u_1^{tr1}), \ldots, (\chi_n^{tr1}, u_n^{tr1}), \ldots, (\chi_N^{tr1}, u_N^{tr1})\}$. This data set is used to train the introduced WSNMF. Cost function (3) is also minimized under the following constraint, which normalizes the columns of W^{tr} to keep the sum of modeled probabilities equal to one after factorization.

$$1 = \sum_{m \in Q_{sz}} (W^{tr})_{mz} \quad \text{for all states } s, \text{ and columns } z \tag{12}$$

where Q_{sz} is the set of elements of the z^{th} columns of W^{tr} which correspond to the s^{th} state.

2) Secondary training

The architecture of the proposed method in the secondary training phase is shown in Figure 2.

During this phase, the procedure of obtaining the GMM weights supervector is repeated for each single speaker of the secondary training data set. Then, the resulting supervector for the k^{th} speaker of the secondary training data set, χ_k^{tr2}, is fed into the NMF trained in the primary phase to estimate its age-gender label $\hat{u}_k^{tr2}=g(\gamma_k^{tr2})$ according to equation (5). In the training phase, the exact chronological age of each speaker is known. Consequently, for the k^{th} estimated age-gender label, the speaker age A_k^{tr2} is known. Therefore, after estimating the age-gender label of all M speakers in the secondary training data set, a secondary set of input-output pairs can be formed such that $S^{tr2}= \{(\hat{u}_1^{tr2}, A_1^{tr2}), \ldots, (\hat{u}_k^{tr2}, A_k^{tr2}), \ldots, (\hat{u}_K^{tr2}, A_K^{tr2})\}$. This set of input-output pairs is used for training the GRNN. We use 10-fold crossvalidation to tune the smoothing parameter of the GRNN Gaussians.

C. Testing phase

Figure 3 indicates the architecture of proposed method in the testing phase. As can be interpreted from this figure, the procedure of obtaining the supervector of the GMM weights is repeated for each single speaker of test data set. Then, the NMF trained in the primary training phase estimates the age-gender label of each supervector. The estimated label clearly shows the gender and age group of a speaker. To estimate the age of a speaker, the estimated age-gender label is fed into the GRNN trained in the secondary training phase. The output of the GRNN is the estimated age of the speaker.

IV. Evaluation and Results

The efficiency of proposed method is assessed on a Dutch corpus. In this section, this corpus is first introduced. Then, evaluation results of the proposed method are presented.

A. Corpora

Speech patterns of 555 speakers from the N-best evaluation corpus [19] were used. The corpus contains live and read commentaries, news, interviews, and reports broadcast in Belgium. Table 1 shows the number of speakers in six different age-gender categories namely Young Male (YM), Young Female (YF), Middle Male (MM), Middle Female (MF), Senior Male (SM), Senior Female (SF).

To evaluate the proposed method, 5-fold cross-validation method is used. Therefore, first all 555 speakers in the database are divided into 5 disjoint folds so that each fold contain 111 speakers. Then, five independent experiments are

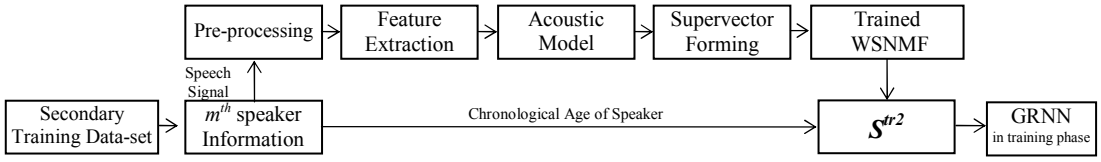

Figure 2. Block diagram of proposed method in secondary training phase

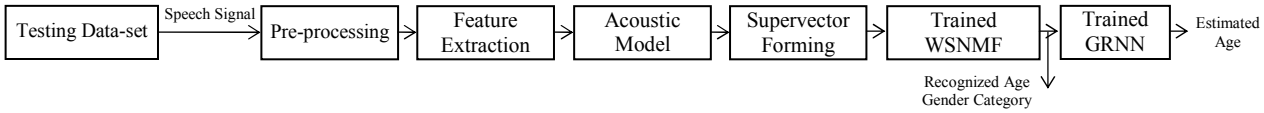

Figure 3. Block diagram of proposed method in testing phase

run so that in each experiment four folds are used as training data set and the rest one fold is used as testing data set. In each of five experiments, 344 out of 444 speakers of the training data set are used as the primary training data set and the rest are used as the secondary training data set.

TABLE I. THE NUMBER OF SPEAKERS IN DIFFERENT AGE-GENDER CATEGORIES

Category Name	**YM**	**YF**	**MM**	**MF**	**SM**	**SF**
Age	18-35	18-35	36-45	36-45	46-81	46-81
Number of Speakers	85	53	160	41	191	25

B. Test results

In all experiments, the number of latent vectors (Z) is 37 and sparsity parameter (ρ) is 1000.

The gender detection accuracy of the proposed method over all five experiments is 96%. Table 2 shows the average of age-group recognition accuracy over all performed experiments. The second row lists the prior class probability, or "chance levels". Hence, the WSNMF method performs better than guessing.

TABLE II. AGE GROUP RECOGNITION ACCURACY IN %

Age Category	**Y**	**M**	**S**
Prior	25	36	39
Recognition Accuracy	38	40	65

Table 3 indicates the relative confusion matrix of proposed method in recognizing six age-gender categories.

TABLE III. THE RELATIVE CONFUSION MATRIX OF PROPOSED METHOD IN RECOGNIZING SIX AGE-GENDER CATEGORIES.

CL / AC	**YM**	**YF**	**MM**	**MF**	**SM**	**SF**
YM	13	03	58	0	26	0
YF	02	77	04	11	057	0
MM	06	01	44	01	47	0
MF	0	54	02	24	17	02
SM	03	01	19	0	76	0
SF	0	2	08	28	28	16

The age estimation accuracy is measured by the mean absolute error, which is calculated as follows.

$$MAE = \frac{\sum_{i=1}^{N} \left| \omega_i - \hat{\omega}_i \right|}{N} \tag{13}$$

where ω_i and $\hat{\omega}_i$ are the real age and estimated age of i^{th} speaker respectively.

The *MAE* of the proposed method over all five experiments is equal to 7.48 years. The *MAE* of taking the average age of all speakers as the estimated age is equal to 8.88 years. This shows that estimating the speaker's age using the proposed approach is 15% better than taking the average true age.

V. CONCLUSIONS

In this paper, a novel hybrid method based on WSNMF and GRNN has been proposed to detect the gender of speakers and estimate their age. In this method, Gaussian mixture weight supervectors of the primary training set are used to train a WSNMF which is applied for recognizing the age-gender category of any unseen speakers. To estimate the age of speakers, a GRNN trained on a secondary training set is inserted in conjunction with trained WSNMF. Evaluation on a Dutch database shows that the *MAE* of age estimation using the proposed hybrid method is 7.48 years. It was also shown that the proposed approach can detect the speaker's gender with 96% accuracy.

REFERENCES

[1] A. K. Jain, P. Flynn, and A. A. Ross, Handbook of biometrics. Springer, 2008.

[2] D. C. Tanner, and M. E. Tanner, Forensic aspects of speech patterns: voice prints, speaker profiling, lie and intoxication detection. Lawyers & Judges Publishing, 2004.

[3] T. Bocklet, A. Maier, J. G. Bauer, F. Burkhardt, and E. Noth, "Age and gender recognition for telephone applications based on GMM supervectors and support vector machines," In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), USA, pp. 1605– 1608, 2008.

[4] F. Metze, et al. "Comparison of four approaches to age and gender recognition for telephone applications," In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), USA, pp. 1089–1092, 2007.

[5] C. Heerden, et al. "Combining regression and classification methods for improving automatic speaker age recognition," In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), USA, pp. 5174-5177, 2010.

[6] M. Feld, F. Burkhardt, and C. Müller, "Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services," In proc. Interspeech, Japan, pp. 2834-2837, 2010.

[7] R. Porat, D. Lange, and Y. Zigel, “Age recognition based on speech signals using weights supervector,” In proc. Interspeech, Japan, pp. 2814-2817, 2010.

[8] X. Zhang, K. Demuynck, and H. Van hamme, “Rapid speaker adaptation with speaker adaptive training and non-negative matrix factorization,” In proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), Czech republic, pp. 4456-4459, 2011.

[9] L. .T Bosch, J. Driesen, H. Van hamme, and L. Boves, “On a computational model for language acquisition: modeling cross-speaker generalization,” In Proc. Int. Conf. Text, Speech and Dialogue, Czech Republic, pp. 315-322, 2009.

[10] Lee, D. D., and H. S. Seung, “Algorithms for non-negative matrix factorization,” Advances in neural information processing systems, pp. 556–562, 2001.

[11] O. Räsänen, and J. Driesen, “A comparison and combination of segmental and fixed-frame signal representations in NMF-based word recognition,” Nordic Conf. Computational Linguistics, NEALT Proceedings Series, vol. 4, pp. 255-262, 2009.

[12] C. Yang, M. Ye, and J. Zhao, “Document clustering based on nonnegative sparse matrix factorization,” In Advances in Natural Computation, Lecture Notes in Computer Science, vol. 3611, pp.557–563, 2005.

[13] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” IEEE Trans. Audio, Speech, and Language Processing, vol. 15, no.3, pp. 1066-1074, 2007.

[14] H. Van hamme, “HAC-models: a novel approach to continuous speech recognition,” In proc. Interspeech, Australia, pp. 2554-2557, 2008.

[15] N. Ho, “Nonnegative matrix factorization algorithms and applications,” PhD thesis, Université. Catholique de Louvain, 2008

[16] D. F. Specht, “A general regression neural network,” IEEE Trans. Neural Networks, vol. 2, no. 6, pp. 568- 576, 1991.

[17] J. Duchateau, M. Wigham, K. Demuynck, and H. Van hamme, “A flexible recogniser architecture in a reading tutor for children,” ITRW on Speech Recognition and Intrinsic Variation, pp. 59–64, 2006.

[18] K. Demuynck, “Extracting, Modelling and Combining Information in Speech Recognition,” Ph.D. thesis, Katholieke Universiteit Leuven, 2001.

[19] D. A. Van Leeuwen, J. Kessens, E. Sanders, and H. van den Heuvel, “Results of the n-best 2008 dutch speech recognition evaluation,” In proc. Interspeech, pp. 2571-2574, 2009.