SPEAKER AGE ESTIMATION USING HIDDEN MARKOV MODEL WEIGHT SUPERVECTORS

Mohamad Hasan Bahari, Hugo Van hamme

Center for Processing Speech and Images (PSI), Department of Electrical Engineering (ESAT), KU Leuven, Belgium
MohamadHasan.Bahari@esat.kuleuven.be, hugo.vanhamme@esat.kuleuven.be

ABSTRACT

This paper proposes a new approach for speaker age estimation. In this method, speakers are modeled by their corresponding Hidden Markov Model (HMM) weight supervectors. Then, Weighted Supervised Non-Negative Matrix Factorization (WSNMF) is applied to reduce the dimension of the input space. Finally, a Least Squares Support Vector Regressor (LS-SVR) is employed to estimate the age of speakers using the obtained low-dimensional vectors. Evaluation results on a corpus of read and spontaneous speech in Dutch confirms the effectiveness of the proposed scheme.

Index Terms— speaker age estimation, weighted supervised non-negative matrix factorization, Least Squares Support Vector Regressor

1. INTRODUCTION

Speech patterns reflect different characteristics of the speaker. 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 [1]. Profiling a person along different characteristics from his/her voice patterns is required in many commercial applications.

In this research, we focus on speaker age estimation. Automated speech-based age estimation is challenging 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 estimation method requires a labeled, wide age-range and balanced database.

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.

Third, voice patterns are affected by many parameters, such as smoking and gender, 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 previously [2,3,14,15]. For example, Bocklet and his colleagues introduced a method based on Support Vector Machines (SVM) to classify speakers into seven age-gender categories [2]. In this method, Gaussian Mixture Model (GMM) mean supervectors were applied as a feature vector for each speaker. In [3], the GMM universal background model is merged with the Support Vector Machine (SVM) classifier and instead of GMM mean supervectors, GMM weight supervectors were used. In [14], Supervised Non-Negative Matrix Factorization (SNMF) was employed to classify speakers based on their age group and genders. In this method, HMM weight supervectors were applied instead of GMM weight or GMM mean supervectors.

Unlike speaker age group recognition [2,3,14,15], the problem of speaker age estimation has not received enough attention in the literature and there are only a few published papers in this field [4,16]. For example, Dobry and his colleagues proposed a speaker age estimator using GMM mean supervectors and a dimension reduction method, namely weighted-pairwise principal components analysis (WPPCA), [16]. In [4], an age estimation and gender recognition approach based on a hybrid architecture of WSNMF and General Regression Neural Networks (GRNN) over HMM weight supervectors was introduced. Although this method could recognize the gender of speakers accurately, its precision in age estimation requires improvement.

In this paper, we introduce a new method to estimate the age of speakers. To develop this method, after determining an HMM acoustic model for all speakers of the database, posterior probabilities of HMM Gaussian mixtures are extracted and concatenated to form a supervector in a high dimensional space for each speaker. We employ WSNMF to reduce its dimensionality. Then, the age of a speaker is estimated using a LS-SVR over the low-dimensional vectors obtained using WSNMF.

This paper is organized as follows. Section 2 introduces WSNMF and LS-SVR. 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.

2. BACKGROUND

In this section, the applied mathematical tools including WSNMF and LS-SVR are briefly reviewed.

2.1. WSNMF

NMF is a popular machine learning algorithm [6], which is successfully applied to different sound and speech processing applications [7-8]. During the last decade, different extensions of NMF such as Supervised NMF (SNMF) [8] and Weighted Supervised NMF (WSNMF) [4] have been developed to solve real world problems. SNMF and WSNMF are originally developed for supervised pattern recognition. However, they can also be applied as dimensionality reduction methods.

2.1.1. WSNMF for pattern recognition

The problem addressed by a WSNMF pattern recognizer 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 features 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 elsewhere. 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 WSNMF directly. First, training data is used to form a matrix V^{tr} as follows.

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

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 $M\times Z$ and $Z\times N$ respectively.

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

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

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

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 [13].

$$\begin{aligned} W^{tr} &\leftarrow \left\{ \frac{\left[W^{tr}\right]}{\left[L(H^{tr})^T\right]} \right\} \circ \left\{ \frac{\left[L \circ V^{tr}\right]}{\left[W^{tr} H^{tr}\right]} (H^{tr})^T \right\} \\ H^{tr} &\leftarrow \left\{ \frac{\left[H^{tr}\right]}{\left[(W^{tr})^T L + \rho\right]} \right\} \circ \left\{ (W^{tr})^T \frac{\left[L \circ V^{tr}\right]}{\left[W^{tr} H^{tr}\right]} \right\} \end{aligned} \tag{4}$$

where $A \circ B$ and $\frac{[A]}{[B]}$ are the element-wise product and division of matrixes A and B respectively, the sign ^{T} is the transpose operator and L is a weighting matrix with the same size as V^{tr}, which is determined as follows.

$$\begin{aligned} L_S &= \beta\left[1_{S\times N}\right] \\ L_B &= \left[1_{B\times N}\right] \end{aligned} \qquad L = \begin{bmatrix} L_S \\ L_B \end{bmatrix} \tag{5}$$

where $1_{S\times N}$ and $1_{B\times N}$ (M=S+B) are two matrices with the same size as V_s^{tr} and V_B^{tr} respectively with all their elements 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

$$\beta = \frac{\sum_{bn} \left(V_B^{tr}\right)_{bn}}{\sum_{sn} \left(V_S^{tr}\right)_{sn}} \tag{6}$$

Calculation of W^{tr} by factorizing the V^{tr} is called training the WSNMF. 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}:

$$\begin{aligned} H^{tst} &= \arg\min_{H^{tst}} D_{KL}\left(x^{tst} \| W_B^{tr} H^{tst}\right) \\ \hat{y} &= g(x^{tst}) = W_S^{tr} H^{tst} \end{aligned} \tag{7}$$

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

2.1.2. WSNMF for dimension reduction

In many real world applications, the dimension of the input space is very high. To reduce this dimension, different methods are suggested such as Principle Component Analysis (PCA), Independent Component Analysis (ICA), and NMF [6,10]. In this paper WSNMF is applied to reduce the dimension of input space. In this approach, training is done on the columns of H^{tr} instead of the columns of V_B^{tr} and testing is performed on H^{tst} instead of x^{tst}. Hence, the dimension of the input space is reduced from M to Z in the training and testing phases. Notice that many other conventional dimensionality reduction approaches, which suffer from high computational complexity, are useless for the current problem because of the large input dimensionality.

2.2. LS-SVR

Least Squares Support Vector Machine (LS-SVM), which is a variant of SVM, was introduced by Suykens and Vandewalle [9]. It is employed as a machine learning tool for regression, clustering and classification tasks. Compared to SVM, LS-SVM enjoys a faster training process because the quadratic programming problem of SVM is reduced to that of solving a system of linear equations. Furthermore, the LS-SVM formulation involves fewer tuning parameters [10].

A continuous function can be fitted to the training data with a Least Squares Support Vector Regressor (LS-SVR), a technique which shares many of the advantages of LS-SVM classification. In this research, LS-SVR is applied to estimate the age of speakers.

3. PROPOSED APPROACH

In this section, the proposed approach for 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 detail.

3.1. Feature selection, acoustic model and supervectors

The acoustic features consist of MEL log-spectra with mean normalization and vocal tract length normalization [5], augmented with their first and second order time derivatives. These features are then mapped to a 36 dimensional acoustic space by means of MIDA, a discriminative linear transformation and decorrelation [11]. 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 (8). 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 independent acoustic model is trained on the CGN corpus (Corpus Gesproken Nederlands) [17].

The speaker dependent HMM weights for each speaker of the Flemish part of the N-Best training corpus [12], result from a re-estimation of the speaker independent weights based on a forced alignment of the training data for that speaker using the speaker-independent acoustic model. Subsequently, the Gaussian mixture weights are extracted and concatenated to form a supervector for each speaker.

For the nth speaker, consider the sth HMM Gaussian mixture with the following state probability density function (PDF).

$$f_S(o_t) = \sum_{j=1}^{J_s} w_j^s \Delta(o_t, \mu_j^s, \Sigma_j^s) \tag{8}$$

where o_t is the acoustic vector at time t, w_j^s is the mixture weight for the jth component of the sth 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 sth mixture. The supervector of Gaussian posterior probabilities can now be formed based on a Estimate-Maximize retraining of the mixture weights on data of speaker n only:

$$\begin{aligned} \lambda^s &= N^s \begin{bmatrix} w_1^s & \cdots & w_q^s & \cdots & w_{Q_s}^s \end{bmatrix}^T \\ \chi_n &= \begin{bmatrix} (\lambda^1)^T & \cdots & (\lambda^s)^T & \cdots & (\lambda^S)^T \end{bmatrix}^T \end{aligned} \tag{9}$$

where χ_n is the supervector of the nth speaker, S is the total number of mixtures, N^s is the number of frames observed in the sth mixture during Viterbi alignment and Q_s is the total number of weights in the sth 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 $(\lambda^S)_q = \sum_t \gamma(q,t)$.

3.2. Training Phase

The block diagram of the training phase of the proposed method is illustrated in figure 1. As in can be interpreted from this figure, the speech signal of each speaker in the training data set is used to form one supervector per speaker. 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. The label of the nth speaker, u_n, which belongs to the dth category, is formed such that the dth element of the label vector is equal to 1 and the other elements are equal to zero. For example, if the jth speaker of the training data set belongs to the second category, its label vector is $u_j^{tr} = [0\ 1\ 0\ 0\ 0\ 0]^T$.

After calculating all N supervectors for all N speakers of the training data set and labeling them with their age-

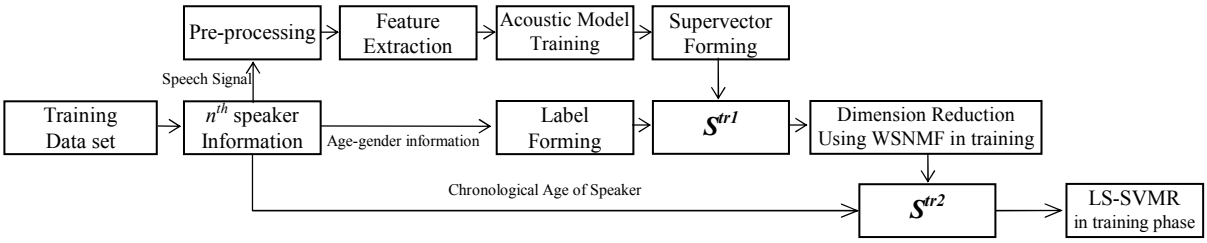

Figure 1. Block diagram of proposed method in training phase

gender category, we obtain $S^{tr1} = \{(\chi_1^{tr}, u_1^{tr}), \ldots, (\chi_n^{t}, u_n^{tr}), \ldots, (\chi_N^{tr}, u_N^{tr})\}$. This data set is utilized to reduce the dimension of input space of the training data set using the introduced WSNMF. In the factorization process, cost function (3) is minimized under the following constraint, which normalizes the columns of W^{tr} such that any linear combination yields valid mixture weights.

$$1 = \sum_{m \in Q_{sz}} (W^{tr})_{mz} \quad \text{for all states } s, \text{ and columns } z \tag{10}$$

where Q_{sz} is the set of elements of the zth column of W^{tr} which correspond to the sth state.

The results of factorization are W^{tr}, and H^{tr}. Now the columns of H^{tr}, which are of size Z, can be used instead of the columns of V_B^{tr}, which are of size M. Consequently, a new input-output set can be formed using the columns of H^{tr} and the choronological age of their corresponding speaker (a_n^{tr}) so that $S^{tr2} = \{(H_1^{tr}, a_1^{tr}), \ldots, (H_n^{tr}, a_n^{tr}), \ldots, (H_N^{tr}, a_N^{tr})\}$. This data set is used to train the LS-SVR with Gaussian kernel function. A 10-fold cross-validation approach is applied to tune the smoothing parameter of the kernels.

3.3. Testing Phase

The block diagram of the proposed method in the testing phase is demonstrated in figure 2. A supervector per test speaker is formed with the same method as outlined in section 3.1. Then, the obtained supervector is fed into the WSNMF, where W^{tr} is fixed and only H^{tst} of size $Z \times 1$ is estimated (7). H^{tst}, which is a low-dimension vector, is used as the input of the trained LS-SVR. The output of the trained LS-SVR is an estimation of the test speaker age.

4. EVALUATION RESULTS AND DISCUSSION

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

4.1. Corpora

Speech patterns of 425 speakers from the Flemish part of the N-best training corpus [12] were used. The corpus contains live and read commentaries, news, interviews, and reports broadcast in Belgium. Figure 3 and 4 show the age histogram of male and female speakers.

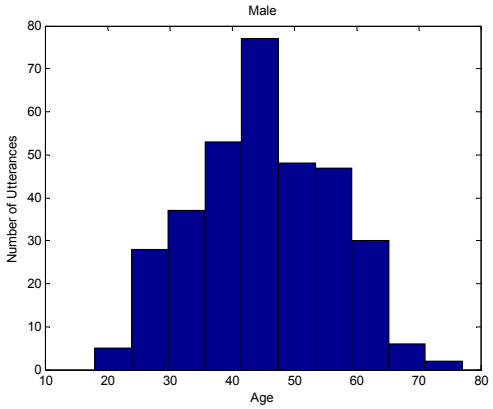

Figure 3. Age histogram of male speakers in the evaluation corpus.

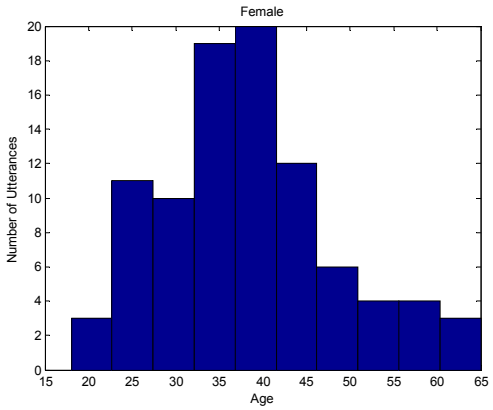

Figure 4. Age histogram of female speakers in the evaluation corpus.

4.2. Test Results

To evaluate the proposed method, 5-fold cross-validation method is used. Therefore, first all 425 speakers in the database are divided into 5 disjoint folds so that each fold contains 85 speakers. Then, five independent experiments are run so that in each experiment, a new fold is used as testing data set and the rest four folds are used as training data set.

The effectiveness of the proposed method is evaluated using the Mean Absolute Error (*MAE*) of the estimated speakers' age, which is calculated as follows.

$$MAE = \frac{\sum_{f=1}^{F} \sum_{i=1}^{N_f} \left| e_{f,i} - \hat{e}_{f,i} \right|}{\sum_{f=1}^{F} N_f} \tag{11}$$

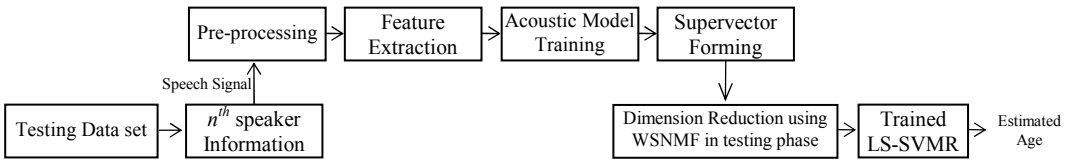

Figure 2. Block diagram of proposed method in testing phase

where F is the total number of independent experiments, which is 5 here. N_f is the number of tested speakers in the fth experiment, which is always equal to 85. $e_{f,i}$ and $\hat{e}_{f,i}$ are the real age and the estimated age of ith speaker in the fth experiment respectively. In all experiments, the sparsity parameter (ρ) is 1000.

The *MAE* of the proposed method for different values of Z is reported in figure 5. For comparison purpose, the results of employing standard NMF [6] as a dimensionality reduction method instead of WSNMF are also included in figure 5 under caption NMF-LSSVMR. As in can be interpreted from this figure, WSNMF is more useful than NMF for dimensionality reduction in this case. This is intuitively due to provided supervision information for WSNMF, while NMF is an unsupervised dimensionality reduction approach.

The *MAE* of our previously published age estimator [4], namely Hybrid WSNMF and GRNN (HWNN), is 8.1. Therefore, the accuracy of the proposed method in speaker age estimation is better than the HWNN when the target dimension Z is between 20 and 70.

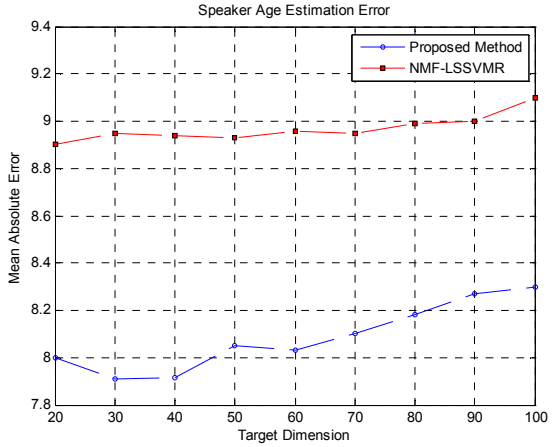

Fig. 5. The MAE of age estimation using the proposed method and NMF-LSSVMR versus target dimension.

5. CONCLUSIONS

In this paper, a novel hybrid method based on WSNMF and LS-SVR has been proposed to estimate the age of speakers. In this method, WSNMF is applied to reduce the dimension of input space, i.e. posterior probability supervectors, and a LS-SVR is used to estimate the age of speakers. Evaluation on a Dutch database confirms the efficiency of the proposed method in speaker age estimation. Furthermore, experiments results show that reducing the dimension of the input space using WSNMF is more effective than with NMF.

REFERENCES

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

[2] 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*), 2008, pp. 1605–1608.

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

[4] M. H. Bahari, and H. Van hamme, "Speaker age estimation and gender recognition based on non-negative matrix factorization," *In proc. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications* (*BIOMS*), 2011.

[5] 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*, 2006, pp. 59–64.

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

[7] 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.

[8] H. Van hamme, "HAC-models: a novel approach to continuous speech recognition," *In proc. Interspeech*, 2008, pp. 2554-2557.

[9] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore, 2002

[10] Fodor, I.K., "Statistical techniques to find similar objects in images," *Proceedings of the American Statistical Association, Statistical Computing Section*, October 2003.

[11] K. Demuynck, *Extracting, Modeling and Combining Information in Speech Recognition*. Ph.D. thesis, Katholieke Universiteit Leuven, 2001.

[12] 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*, 2009, pp. 2571-2574.

[13] N. Ho, *Nonnegative matrix factorization algorithms and applications*. PhD thesis, Université. Catholique de Louvain, 2008.

[14] M. H. Bahari and H. Van hamme, "Age and Gender Recognition from Speech Patterns Based on Supervised Non-Negative Matrix Factorization," *In Proc. IAFPA 20th Annual Conference*, 2011, pp 3-5.

[15] 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

[16] G. Dobry, et. al. "Supervector Dimension Reduction for Efficient Speaker Age Estimation Based on the Acoustic Speech Signal," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no.v, pp. 1975-1985, 2011.

[17] N. H. J. Oostdijk and D. Broeder, "The spoken Dutch corpus and its exploitation environment" in proc. 4th Int. Workshop on Linguistically Interpreted Corpora, Budapest, Hungary, 2003.