Factor Analysis based Session Variability Compensation for Automatic Speech Recognition

Mickael Rouvier, Mohamed Bouallegue, Driss Matrouf, Georges Linarès

University of Avignon, LIA, France
{mickael.rouvier, driss.matrouf, georges.linares}@univ-avignon.fr
mohamed.bouallegue@etd.univ-avignon.fr

***Abstract*—In this paper we propose a new feature normalization based on Factor Analysis (FA) for the problem of acoustic variability in Automatic Speech Recognition (ASR). The FA paradigm was previously used in the field of ASR, in order to model the usefull information: the HMM state dependent acoustic information. In this paper, we propose to use the FA paradigm to model the useless information (speaker- or channel-variability) in order to remove it from acoustic data frames. The transformed training data frames are then used to train new HMM models using the standard training algorithm. The transformation is also applied to the test data before the decoding process. With this approach we obtain, on french broadcast news, an absolute WER reduction of 1.3%.**

I. Introduction

The goal of Automatic Speech Recognition (ASR) is to extract the linguistic information from a given speech signal recording. However, the speech signal includes not only the linguistic information, but also some disturbing information [1]. The disturbing information covers a large panel of variabilities such as speaker variability (vocal tract length [2], speaker spontaneity [3]...), recording conditions (background noise [4], microphone setup and transmission channel). The speech signal we observe is then composed of useful information (the linguistic information) but also useless information, named here the session variability. The variability of the channel, speaker and environment are one of the most important factors affecting the performance of the ASR. We need to find a way to model this useless information in order to remove it.

Previously, a number of methods for reducing these variability effects were proposed in the feature domain. Cepstral Mean Subtraction (CMS) [5] is applied to remove linear channel variability. An extension of this approach is proposed and involved normalizing the distribution of single cepstral features (over some specific window length) by subtracting their mean and dividing by their standard deviation. RASTA processing has been shown to improve the recognition performance in presence of convolutional distortions and additive noise [6]. Another interesting normalization to counter speaker variability and that affects the features is the Vocal Tract Length Normalization (VTLN) [2]. VTLN attempts to normalize speech representation by removing the differences caused by the variance in the length of the speaker's vocal tract.

Recently a Factor Analysis (FA) approach was applied in the speaker recognition domain to model the session variability as additive component [7]. The basic idea behind this approach is that the session component is located in low-dimensional acoustic subspace. In this manner the FA paradigm allows a balance between the amount of parameters to be estimated and the amount of training data. In fact the session component is decomposed into two parts : 1- the session variability subspace basis (consisting to estimate large number of parameters) but estimated on the whole training data, 2- and the session coordinates in this subspace (a small number of factors) which are estimated on a single recording.

Some authors have proposed to apply the FA paradigm in ASR system. These investigations were carried out on the accurate modeling of useful information (Subspace Gaussian Mixture Model (SGMM) [8], [10] and Canonical State Models (CSM) [9]), but not on the nuisance variability modeling. We propose, in this paper, the use of FA paradigm to model the session variability component and hence to remove it directly from the speech data frames. Another contribution of this paper is to extend the FA paradigm in order to deal with multiple variabilities in the speech signal.

The paper is organized as follows: in the next section, we explain how to model the variability effect within the FA paradigm. In Section 3, we propose an extended FA paradigm that deals with multiple variability. Results are reported and commented in Section 4. Finally, we conclude and we present some perspectives in Section 5.

II. Session variability modeling

The speech signal conveys not only linguistic information but also useless information. This useless information can be of very different nature and can be related to environment-variability (background noise...), speaker-variability (gender, age, emotion...), channel-variability (microphone...)... This useless informations is present in the speech signal and affects the Hidden Markov Models (HMM) of an ASR systems. In order to model only linguistic information in the HMM, a solution could be to remove the useless information from the speech data frames.

The Factor Analysis (FA) paradigm gives the possibility to model useless information (in a subspace of low dimension R) in order to remove it from the speech data frames. Let G be a set of Gaussians structuring the acoustic space of the speech signal (we will call this normalization acoustic model). Let m be the supervector obtained by the concatenation of all means in G. Let i the useful information to be modeled and h be

the session information (that represents speaker- or channel-variability). By using the FA paradigm, the supervector $m_{i,h}$ (random variable) can be decomposed into three different components:

$$m_{i,h} = m + Dy_i + Ux_h \quad (1)$$

Here m is the supervector composed of the Gaussian means coming from set G. The set of Gaussians G is trained on a large amount of data containing useful and useless information. m is a vector of dimension MD, where M is the number of Gaussian in the GMM G and D the dimension of acoustic space. y_i models the useful information, it's a vector of dimension MD. It can correspond to the linguistic content of a given recording, to a phoneme or to an HMM state. Ux is the session variability component. U is composed by the eigenvectors of the session variability. x_h are the session factors, it's a vector of dimension R. Both y_i and x_h are assumed to be normally distributed among $N(0, I)$. D is a diagonal matrix (a $MD \times MD$ matrix) so that DD^t is the *a priori* covariance matrix of the useful component. U is a rectangular matrix (a $MD \times R$ matrix) so that UU^t is the *a priori* covariance matrix of the session component random vector.

As shown in Equation 1, the success of FA modeling depends mainly on the assumption that the nuisance variability is located in a subspace of low dimension (dimension R) and on assumption that the session effect is additive.

In order to have a balance between the modeling precision and the amount of the data leading to accurate parameters estimation, we have chosen i to be a context-independent phoneme. In fact, if we take i as being a part of a phoneme, for example HMM state, for several states we can not have a sufficient number of frames to estimate the session factors x_h. In this section we will consider the speaker and the channel variabilities as a session. By taking i as a context-independent phoneme, the model of Equation 1 can be written more explicitely as follows:

$$m_{phoneme,session} = m + Dy_{phoneme} + Ux_{session} \quad (2)$$

The U matrix is global and common to all phonemes. It is estimated by using a large corpus of phonemes produced by several speakers and under a variety of acoustic conditions. In this manner we can isolate the session variability (speaker and channel). It is important to note that the model of Equation 2 is not used for speech recognition, but only to compensate speech data frames. The compensated data frames can then be used for training or testing an ASR.

A. Discussion about the normalization acoustic model G

FA parameter estimation (fully described in [11]) are based on the following zeroth and first order statistics:

$$N_g^{(h,s)} = \sum_{t \in (h,s)} \gamma_g(t) \quad ; \quad X_g^{(h,s)} = \sum_{t \in (h,s)} \gamma_g(t) \cdot t \quad (3)$$

where g is a subscript of a Gaussian in G. $\gamma_g(t)$ is the *a posteriori* probability of Gaussian g given the cepstral vector of observation t. We can see that the accuracy of the FA parameter estimation is mainly based on the *a posteriori* probabilities $\gamma_g(t)$. The estimation of these probabilities depend on the used acoustic model and the associated decoding procedure. Maybe, more robust manner to obtain theses probabilities could be to use a complete ASR. In fact, by using an ASR, we not only use acoustic information, but also linguistic information contained in the lexical and the language models. In this case the supervector m of Equation 1 is the concatenation of all Gaussian means contained in the HMM : G is the set of all the Gaussians in the HMM. In this preliminary study we have used a GMM-UBM instead of an HMM. Which probably tends to less accurate *a posteriori* probability estimations. In this case the mean supervector is the concatenation of all Gaussian means contained in the GMM-UBM (G is composed by the Gaussians in the UBM). This GMM-UBM is trained by using data coming from a large number of speakers and different channel sources.

B. Estimating the session-variability subspace

The U matrix is a global parameter. It is estimated using a large amount of data containing session variability. The matrix is iteratively estimated using the Expectation Maximization (EM) algorithm. Each step, $x_{session}$ (session factors) are estimated, then $y_{phoneme}$ is estimated for each phoneme (using the new x) and finally U is estimated globally, based on these $x_{session}$ and $y_{phoneme}$. Since $x_{session}$ and $y_{phoneme}$ also depend on U, the process is iterated. The step by step algorithm is described in [11].

C. ASR acoustic model training

Each utterance in the training corpus is first normalized with respect to the session variability using the following equation:

$$\hat{t} = t - \sum_{g=1}^{M} \gamma_g(t) \cdot \{U \cdot x_{utterance}\}_{[g]} \quad (4)$$

where M is the number of Gaussian components in the UBM, $\gamma_g(t)$ is the a posteriori probability of Gaussian g given by frame t. These probabilities are estimated by using the UBM. And $U \cdot x_{utterance}$ is the session variability component estimated on the utterance recording. It's a supervector with $M \times D$ components. $\{U \cdot x_{utterance}\}_{[g]}$ is the g^{th} D component bloc vector of $U \cdot x_{utterance}$.

After normalizing all utterances using Equation 4, HMMs for ASR are trained using the normalized speech data frames. Theoretically, for each utterance we must estimate the session variability component on each phoneme and normalize it with the specific session variability component. In practice, this is not feasible because each phoneme in an utterance contain too few frames. So we estimate the session variability component globally on the utterance and normalize the features.

III. Modeling multiple session variability

In previous section, in Equation 2, U matrix models a specific kind of session variability (speaker- or channel-variability). But the variabilities in ASR are multiple [1]. We propose a modified version of FA in order to deal with multiple-variability effects. We extend the FA paradigm, for that each matrix can model a specific variability. Here, the matrix U models speaker-variability and the matrix V models channel-variability. The modified FA version is written:

$$m_{observed} = m_{ubm} + Dy_{phoneme} + Ux_{speaker} + Vz_{channel} \tag{5}$$

where, as previously, m is a mean supervector, y is the part which is specific to the context-independent phoneme, weighted by D. But in this section Ux is the speaker-variability component and Vz is the channel variability component.

Previously, an estimate of U matrix is obtained from one corpus where all session contain a specific variability. Here, an estimate of each matrix is obtained from a different corpus, that allows to estimate the matrix U and V on specific variability. The U matrix is estimated on the speaker-variability corpus, where each session represent phoneme-speaker. And the V matrix is estimated on the channel variability corpus, where each session represents phoneme-channel pair.

In Speaker Verification (SV), the authors propose a framework [12] similar to the Equation 5. The framework called Joint Factor Analysis (JFA) model on the same corpus, two session variability refer as speaker and channel factors. However, the framwork that we propose it's a variant of JFA. The session variability are estimated iteratively and we propose to modelise the session variability on two different corpus. This will allow to extend the framework to other variability.

A. Estimating speaker- and channel-variability spaces

The matrices U and V are common to all phonemes, they are jointly optimized. The estimation procedure is presented in Algorithm 1. In the first step, the U matrix is optimized on the data speaker corpus. The $x_{speaker}$ and $z_{channel}$ vectors are estimated, then $y_{phoneme}$ is estimated for each phoneme (using the new x and z) and finally U is estimated globally, based on these x, z and y. In a second step, the V matrix is optimized on the data channel corpus. The $x_{speaker}$ and $z_{channel}$ vectors are estimated, then $y_{phoneme}$ is estimated for each phoneme (using the new x and z) and finally V is estimated globally, based on these x, z and y.

The center statistics are calculated to take into account U and V matrices:

Algorithm 1: Estimation algorithm for U and V

For each phoneme s and session h: $y_s \leftarrow 0$, $x_{(h,s)} \leftarrow 0$, $z_{(h,s)} \leftarrow 0$;
$U \leftarrow random$ (U is initialized randomly);
$V \leftarrow random$ (V is initialized randomly);
Estimate statistics: N^s, $N^{(h,s)}$, X^s, $X^{(h,s)}$ on speaker variability corpus;
Estimate statistics: M^s, $M^{(h,s)}$, Z^s, $Z^{(h,s)}$ on channel variability corpus;
for $i = 1$ *to* $nb\_iterations$ **do**
  for *all* h *and* s *of the speaker variability corpus* **do**
    Center statistics: $\overline{Z}^{(h,s)}$;
    Center statistics: $\overline{X}^{(h,s)}$;
    Estimate $L^{-1}_{(h,s)}$ and $B_{(h,s)}$;
    Estimate $z_{(h,s)}$;
    Estimate $x_{(h,s)}$;
    Center statistics: $\overline{Z}^s$;
    Center statistics: $\overline{X}^s$;
    Estimate y_s;
  end
  Estimate matrix U ;
  for *all* h *and* s *of the channel variability corpus* **do**
    Center statistics: $\overline{Z}^{(h,s)}$;
    Center statistics: $\overline{X}^{(h,s)}$;
    Estimate $P^{-1}_{(h,s)}$ and $Q_{(h,s)}$;
    Estimate $z_{(h,s)}$;
    Estimate $x_{(h,s)}$;
    Center statistics: $\overline{Z}^s$;
    Center statistics: $\overline{X}^s$;
    Estimate y_s;
  end
  Estimate matrix V ;
end

$$\begin{aligned}
\overline{X}^s_g &= X^s_g - \sum_{h \in s} N^{(h,s)}_g \cdot \{m + Ux_{(h,s)} + Vz_{(h,s)}\}_g \\
\overline{X}^{(h,s)}_g &= X^{(h,s)}_g - N^{(h,s)}_g \cdot \{m + Dy_s + Vz_s\}_g \\
\overline{Z}^s_g &= Z^s_g - \sum_{h \in s} M^{(h,s)}_g \cdot \{m + Ux_{(h,s)} + Vz_{(h,s)}\}_g \\
\overline{Z}^{(h,s)}_g &= Z^{(h,s)}_g - M^{(h,s)}_g \cdot \{m + Dy_s + Ux_s\}_g
\end{aligned} \tag{6}$$

where N^s, $N^{(h,s)}$, X^s, $X^{(h,s)}$ are the zeroth- and first-order statistics, calculated on the speaker variability corpus and M^s, $M^{(h,s)}$, Z^s, $Z^{(h,s)}$ are the zeroth- and first-order statistics, calculated from channel variability corpus.

Let $L_{(h,s)}$ and $P_{(h,s)}$ be a $R \times R$ dimensional matrix and $B_{(h,s)}$ and $Q_{(h,s)}$ a vector for dimension R, defined by:

$$
\begin{aligned}
B_{(h,s)} &= \sum_{g \in UBM} U_g^T \cdot \Sigma_g^{-1} \cdot \overline{X_g^{h,s}} \\
L_{(h,s)} &= I + \sum_{g \in UBM} N_g^{(h,s)} \cdot U_g^T \cdot \Sigma_g^{-1} \cdot U_g \\
Q_{(h,s)} &= \sum_{g \in UBM} V_g^T \cdot \Sigma_g^{-1} \cdot \overline{Z_g^{h,s}} \\
P_{(h,s)} &= I + \sum_{g \in UBM} M_g^{(h,s)} \cdot V_g^T \cdot \Sigma_g^{-1} \cdot V_g
\end{aligned}
\tag{7}
$$

where Σ_g is the covariance matrix of the g^{th} UBM component. By using $L_{(h,s)}$, $B_{(h,s)}$, $P_{(h,s)}$ and $Q_{(h,s)}$ we can obtain $x_{h,s}$, $z_{h,s}$ and y_s from the following equations :

$$
\begin{aligned}
z_{h,s} &= P_{(h,s)}^{-1} \cdot Q_{(h,s)} \\
x_{h,s} &= L_{(h,s)}^{-1} \cdot B_{(h,s)} \\
y_s &= \frac{\tau}{\tau + N_g} \cdot D_g \Sigma_g^{-1} \cdot \overline{X_g^{h,s}}
\end{aligned}
\tag{8}
$$

where $D_g = (1/\sqrt{\tau})\Sigma_g^{1/2}$ (τ is set to 14.0 in our experiments).

Finally the U and V matrix can be estimated row by row, with U_g^i and V_g^i being the i^{th} row of U_g and V_g; thus :

$$
\begin{aligned}
U_g^i &= \mathcal{L}(g)^{-1} \cdot \mathcal{R}^i(g) \\
V_g^i &= \mathcal{P}(g)^{-1} \cdot \mathcal{Q}^i(g)
\end{aligned}
\tag{9}
$$

where where $\mathcal{L}(g)$, $\mathcal{R}^i(g)$, $\mathcal{P}(g)$ and $\mathcal{Q}^i(g)$ are given by:

$$
\begin{aligned}
\mathcal{L}(g) &= \sum_s \sum_{h \in s} N_g^{(h,s)} \cdot (L_{(h,s)}^{-1} + \mathbf{x}_{(h,s)} \mathbf{x}_{(h,s)}^T) \\
\mathcal{R}^i(g) &= \sum_s \sum_{h \in s} \overline{X}_g^{(h,s)}[i] \cdot \mathbf{x}_{(h,s)} \\
\mathcal{P}(g) &= \sum_s \sum_{h \in s} M_g^{(h,s)} \cdot (P_{(h,s)}^{-1} + \mathbf{x}_{(h,s)} \mathbf{x}_{(h,s)}^T) \\
\mathcal{Q}^i(g) &= \sum_s \sum_{h \in s} \overline{Z}_g^{(h,s)}[i] \cdot \mathbf{x}_{(h,s)}
\end{aligned}
\tag{10}
$$

B. ASR acoustic model training on multiple variability

Once the U and V matrices are obtained, features are normalized in order to remove speaker and channel effects on this features. As previously, the adaptation of each vector is obtained by subtracting from the observation feature the speaker- and channel-variability component:

$$
\hat{t} = t - \sum_{g=1}^{M} \gamma_g(t) \cdot (\{U \cdot x_{utterance}\}_{[g]} + \{V \cdot z_{utterance}\}_{[g]})
\tag{11}
$$

where $U \cdot x_{utterance}$ and $V \cdot z_{utterance}$ are the channel components estimated on the recording. The latent variable $x_{utterance}$, $z_{utterance}$ are estimated following the Algorithm 1 and setting the U and V matrix.

As previously, after normalizing all utterances using Equation 11, HMMs for speech recognition are trained by using the normalized speech data frames.

IV. System Description and Results

A. Speeral

For this test, we used the LIA broadcast news ASR system, SPEERAL [13]. This system is based on an A* decoder using state-dependent HMM for acoustic modeling. The baseline Language Model (LM) is a 65k word broadcast news 3-gram, estimated on 200M words from the French newspaper "Le Monde" and from the ESTER broadcast news training corpus of about 1M words. The system uses context-dependent models trained on the 90 hours of ESTER transcribed data. State tying is performed by a decision tree algorithm, using acoustic context related questions.

B. System and Corpus

The speech transcription process is composed of two passes:

1) The first pass (*PASS-1*) uses the acoustic model corresponding to the gender and the bandwidth detected by the segmentation process, and using a trigram language model.
2) The second pass (*PASS-2*) applies a MLLR transformation by speaker or by segment, and uses the same trigram language models as the first pass.

The performance of the system is evaluated on the ESTER evaluation corpus. The data was extracted from French radio broadcast news. It is composed of 18 audio files, with a total duration of 10h.

C. ASR acoustic model training

The acoustic model was learned on a training corpus where all data frames are normalised by Equations 4 or 11. The normalization is also applied to test data frames to be decoded. For all these results the rank of U and V matrices is fixed to 60. The GMM-UBM used in the FA approach is composed of 600 Gaussians.

D. ASR acoustic model training on a specific variability

In a first step, we compare the results of our baseline with the systems training on a specific variability. *Norm-speaker* and *Norm-channel* are the systems where acoustic models are trained on a specific variability using Equation 2.

TABLE I
Results in % WER on ESTER corpus dealing with specific variabilities

	PASS-1	PASS-2
Baseline	29.6	27.5
Norm-speaker	28.5	26.9
Norm-channel	28.6	26.7

Table I shows the results obtained on ESTER corpus. In *PASS-1*, we observe that baseline obtained a Word Error

Rate (WER) of 29.6% and that *Norm-speaker* and *Norm-channel* obtained a WER of 28.5% and 28.6% respectively (an absolute WER improvements respectively of 1.1% and 1.0%). In *PASS-2*, we obtained an absolute WER for Norm-channel and Norm-speaker respectively of 0.8% and 0.6%. If the gains are less important than in PASS-1, this may be due to the MLLR adaptation. Indeed, the MLLR technique adapt acoustic model to a particular speaker in capturing general relationships between the original model and the current speaker or new acoustic environment. The new speaker dependent models allow to reduce the intra-speaker-variability.

E. ASR acoustic model training on multiple variabilities

Table II shows the results using the extended FA paradigm. *Norm-speaker-channel* is the system whose acoustic models are train on a multiple variability using Equation 5. Compared to our baseline, *Norm-speaker-channel* system obtained in *PASS-2*, an absolute WER gain of 1.3%. In previous section the best system (*Norm-channel*) obtained in *PASS-2* an absolute WER improvement of 0.8%.

These results confirm that the extended FA paradigm must model different nuisance variability and remove it in the acoustic space. In these experiments we limit to the speaker and channel variability. But it could be possible in this extended FA paradigm to remove other nuisance variability.

TABLE II
RESULTS IN % WER ON ESTER CORPUS DEALING WITH MULTIPLE VARIABILITIES

	PASS-1	PASS-2
Baseline	29.6	27.5
Norm-speaker-channel	28.0	26.2

In Table III, we observe the *Norm-speaker-channel* system robustness by evaluating sentences. We have sorted the sentences of the baseline system into 11 ranges of WER. Each sentence of the *Norm-speaker-channel* system is put in the same range as the sentence of the baseline. The difference between the two systems is mainly based on the normalization of the data frames. This Table allows to compare sentences according to their difficult acoustic conditions.

TABLE III
RESULTS FOR EACH % WER RANGE

WER Range %	Baseline	Norm-spk-cha	Absolute WER reduction
0-5	0.35	2.46	-2.11
5-10	7.19	8.52	-1.33
10-15	12.73	13.44	-0.70
15-20	17.60	18.36	-0.75
20-25	21.52	20.91	0.61
25-30	26.71	25.80	0.91
30-35	32.18	29.79	2.39
35-40	37.01	35.79	1.22
40-45	41.85	39.45	2.39
45-50	46.36	44.00	2.36
50-100	68.28	62.54	5.74

We can observe, for the 0-10 range, that WER between *Norm-speaker-channel* and *Baseline* is increased. We obtain on the 0-5 range an absolute WER increase of 2.11%. On the 20%-100% range, we observe some gains on *Norm-speaker-channel* systems. This gain is particularly important on the 50%-100% range (an absolute WER reduction of 5.74%). More over, if we observe on ESTER corpus, an absolute WER reduction of 1.3%. The normalization is mainly interesting on sentences with difficult acoustic conditions. However, on sentence with a low WER the normalization gives no improvement. The reason may be that the speaker- and channel-variabilities are mainly concentrated on sentences with high WER.

V. CONCLUSION

In this work, we proposed a framework of feature normalization based on the FA paradigm. We also presented an extension to deal with multiple and different types of variabilities. This extension can be used for other variabilities which can be studied in the future. In this preliminary study, we have used a GMM-UBM instead of HMM for the normalization of acoustic model G. We try in next work to see if replacing the model, by an HMM, may improve performance.

REFERENCES

[1] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10-11, pp. 763 – 786, 2007, intrinsic Speech Variations.

[2] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, pp. 346–348, 1996.

[3] R. Dufour, F. Bougares, Y. Estve, and P. Delglise, "Unsupervised model adaptation on targeted speech segments for lvcsr system combination," in *Interspeech 2010*, Makuhari (Japan), 26-30 september 2010.

[4] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition," *Speech Communication*, vol. 45, no. 4, pp. 401 – 423, 2005.

[5] M. Westphal, "The use of cepstral means in conversational speech recognition," in *In Proceedings of the European Conference on Speech Communication and Technology (Eurospeech*, 1997, pp. 1143–1146.

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," in *IEEE Transactions on Speech and Acoustics*, vol. 2, October 1994, pp. 587–589.

[7] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345 – 354, may 2005.

[8] M. Bouallegue, D. Matrouf, and G. Linares, "A simplified subspace gaussian mixture to compact acoustic models for speech recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, 2011.

[9] M. Gales and K. Yu, "Canonical state models for automatic speech recognition," in *Interspeech 2010*, Makuhari (Japan), 26-30 september 2010.

[10] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafi andt, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "Subspace gaussian mixture models for speech recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 4330 –4333.

[11] D. Matrouf, N. Scheffer, B. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *Interspeech 2007*, 2007.

[12] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," CRIM, Tech. Rep., January 2006.

[13] G. Linarès, P. Nocera, D. Massonié, and D. Matrouf, "The lia speech recognition system: from 10xrt to 1xrt," in *Proceedings of the 10th international conference on Text, speech and dialogue*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 302–308.