FACTOR ANALYSIS-BASED INFORMATION INTEGRATION FOR ARABIC DIALECT IDENTIFICATION

Yun Lei and John H.L. Hansen

CRSS: Center for Robust Speech Systems
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas, Richardson, Texas 75083,USA
{yxl059200, John.Hansen}@utdallas.edu

ABSTRACT

In this study, we propose a new factor analysis-based modeling technique to more clearly describe the composition of the supervector defined by the GMM model for dialect identification. The method utilizes knowledge types of information contained in the transcript file of the data. We evaluate the effects of the proposed modeling algorithm on a GMM-based Arabic dialect identification system. In particular, we compare eigenchannel modeling and our proposed information integration modeling. We show that the proposed modeling can obtain a 4.23% relative EER reduction with the same total number of factors, and a 9.37% relative EER reduction with the same number of channel/session factors versus eigenchannel modeling.

Index Terms— Arabic, dialect identification, factor analysis, information integration

1. INTRODUCTION

Dialect identification is an emerging research topic in the speech recognition community because dialect is one of the most important factors next to gender that influences speech recognition performance. Automatic dialect identification or classification is important for characterizing speaker traits and knowledge estimation which can be used in many fields. The definition for *dialect* in this study is a pattern of pronunciation and/or vocabulary of a language used by the community of native speakers belonging to some geographical region. In previous studies, a Gaussian Mixture Model (GMM) based classifier has been applied for unconstrained data [1]. There are also successful methods based on reducing model confusion for improved performance for dialect classification [2, 3, 4].

Factor analysis has proven to be effective for speaker recognition [5, 6], language recognition [7] and dialect identification [8]. Eigenchannel modeling, is an approach for channel compensation in the model domain [9] or feature domain [10], and is employed to characterize distortions of the session/utterance via a small number of parameters in a lower dimensional subspace, called the channel factors. Eigenvoice modeling, as an approach for speaker adaptation, greatly reduces the number of parameters (e.g., speaker factors) to be estimated for the new speaker.

In this study, a new modeling approach called information integration based on factor analysis (IIFA) is described. IIFA modeling is applied for analyzing the composition of the supervector of the utterance by integrating multiple types of information contained in the audio stream knowledge[1]. Application of IIFA modeling for dialect identification can significantly improve dialect ID performance.

2. REVIEW OF FACTOR ANALYSIS

The Gaussian Mixture Model has been a standard approach in speaker and language identification. Factor analysis, as an adaptation model, has been successfully applied for GMM-based systems to address mismatch. In speaker recognition, the supervector obtained by concatenating all mean vectors in the GMM corresponding to a given utterance is applied as the representation of the utterance. Since the supervector M is speaker- and channel-dependent, it can be decomposed into a sum of two supervectors, a speaker supervector s and a channel/session supervector c:

$$M = s + c, \tag{1}$$

where s and c are statistically independent. Furthermore, the speaker supervector s can be decomposed into a sum of two parts, a speaker- and channel-independent supervector m which is the supervector of a UBM and a speaker-dependent

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029, and the University of Texas at Dallas under Project EMMITT. Approved for public release; distribution unlimited.

[1] In this study, the types of knowledge which are incorporated into the proposed IIFA scheme include: gender, speaker, topic, dialect, session/channel. We assume this information is part of the audio stream knowledge and that it is contained in the corresponding transcription. However, all evaluations here employ unsupervised train/test data, so no text transcripts are ever used, only the corresponding audio stream knowledge.

supervector m_s:

$$s = m + m_s. \tag{2}$$

It is assumed that the distribution of m_s and c can be described by some hidden variables:

$$m_s = v \cdot y, \tag{3}$$

$$c = u \cdot x, \tag{4}$$

where v and u are the rectangular matrixes of low rank, and y and x are normally distributed random vectors. The columns of v and u are referred to as eigenvoices and eigenchannels; and the components of y and x are referred to as speaker and channel factors. Based on the decomposition above, the supervector M of the given utterance can be rewritten as:

$$M = m + m_s + c = m + v \cdot y + u \cdot x. \tag{5}$$

In speaker recognition processing, the speaker factors are employed to describe the speaker traits with limited data; the channel factors are employed for channel composition against the mismatch between training and test data.

In language or dialect recognition, similar models can be employed. Since the amount of adaptation data is generally large enough to perform classic MAP adaptation, which actually is a special case of factor analysis, from the UBM, it is not necessary to employ a model such as eigenvoice to decrease the dimensions of the variables (e.g., the vector y in Eq.5), and therefore only channel/session composition is considered. The supervector M of the given utterance can be represented as:

$$M = m + m_d + c, \tag{6}$$

where m is the supervector of the UBM, m_d is the language/dialect dependent supervector which is obtained from MAP adaptation, and c is the channel supervector which corresponds to the same c in Eq. 5 for speaker recognition.

3. INFORMATION INTEGRATION BASED ON FACTOR ANALYSIS: IIFA

Factor analysis has proven to be effective for GMM speaker recognition and language/dialect recognition. A large data pool is typically used to estimate the eigenspace v and u using PCA or EM iterations. In speaker recognition, the speaker information in the transcription of the data pool is used to estimate the eigenvoices, and session information in the transcription of the data pool is used to estimate the eigenchannels (Note, in general we assume each utterance represents one session). In language/dialect recognition, a similar procedure is applied for estimation of the eigenchannels. However, if we consider the used data pool carefully, many types of information can be contained in the transcription data, such as gender, age, language, dialect, accent, speaker, channel, topic, and so on. Normally, only one type of information from the transcripts is used for special problems such as speaker information for speaker recognition. Here, it is suggested that if more information is appended in solving the problem, better performance can be obtained. The proposed Information Integration based on Factor Analysis (IIFA) algorithm integrates all information from the audio stream knowledge into the identification system. In the IIFA model, for each session/utterance, the supervector of the session can be expressed as:

$$M = m + \sum_{k=1}^{N} m_k, \tag{7}$$

where M is the supervector of the session, m is the supervector of the UBM, m_k is the k^{th} supervector of the session, and N is the number of the factor components used in the model. Here, m_k is the *factor component* representing the influence from the k^{th} factor type for the session such as supervector m_s and c in Sec.2. All factor components m_k are assumed to be statistically independent and N is dependent on the information which can be found in the transcript or other sources. Furthermore, it is assumed that the distribution of the supervector m_k can be described by some variables:

$$m_k = u_k \cdot x_k, \tag{8}$$

where u_k is referred to as the eigenspace of the k^{th} factor component, and x_k is referred to as the random vector of the k^{th} factor component. Under this model, when $N = 1$, the IIFA model degenerates to the eigenchannel or eigenvoice model individually; when $N = 2$, the IIFA model is equal to a mixture of eigenchannel and eigenvoice models.

In the IIFA model, all factor components m_k can be classified into three classes based on different problems: positive components, negative components, and neutral components. Positive components should support problem solving; negative components are supposed to reduce problem solving; and neutral components have little impact on solving the problem. To improve the performance of identification, all positive components should be kept and described carefully, all negative components should be suppressed as much as possible, and neutral components can be set aside since they do not have significant impact on identification performance. For different problems, the class of the factor component m_k may be different. For example, in speaker recognition, the m_k representing the speaker is the positive component, but it becomes the negative component in language identification. In a manner similar to estimation of u and v in eigenchannel and eigenvoice models, PCA without iterations or EM training can be used to compute the eigenspace u_k; the factors x_k can be estimated for each test utterance in a step-by-step manner using a similar algorithm as that used for eigenchannel and eigenvoice models [5, 11].

4. SYSTEM DESCRIPTION

In this section, the GMM-based dialect recognition system is described. The IIFA models are applied for the system.

4.1. Arabic dialect corpus

Since the design and implemention of the IIFA model are dependent on knowledge in the speech data's transcripts, the dialect corpus used in this study is described first. The corpus employed for this study consists of Arabic dialect data from 5 different regions, including United Arab Emirates (UAE), Egypt, Iraq, Palestine, and Syria. A full set of 250 sessions (500 speakers) are recorded in the corpus, with 100 speakers per dialect. A set of 13 pre-selected topics were chosen with the aim of achieving as much as possible an equal distribution across all topics for the final database. The dialect, gender, speaker and topic are labeled per recording per conversation. After a silence removal process, the training data includes 43 hours of speech from the 5 dialects from 340 speakers; the testing data consists of 20 hours of speech including 598 true trials and 2392 impostor trials, with about 2 minutes per trial.

4.2. IIFA system

Since gender, dialect, speaker, and topic are labeled in the transcript, the IIFA model in the dialect identificatin system is designed as:

$$M = m + m_d + m_g + m_s + m_t + m_c, \qquad (9)$$

where M is the supervector of the session, m is the supervector of the UBM, m_d is the factor component representing the dialect, m_g is the factor component representing the gender, m_s is the factor component representing the speaker, m_t is the factor component representing the topic, and m_c is the factor component representing the session. The number of factor components in the IIFA model N is equal to 5. Since the purpose of the system is to identify the dialect, m_d will be a positive component while m_g, m_s and m_c are assigned as negative components. The topic can be considered a neutral component and can be set aside in the GMM based system since topic is a factor related with the language domain, not the acoustic domain. The UBM is trained on the entire training data, and is adapted to dialect-dependent GMM models using classic MAP adaptation where m_d can be obtained. All negative components are described by the product of the eigenspace and the corresponding vector as:

$$m_g = u_g \cdot x_g, \qquad m_s = u_s \cdot x_s, \qquad m_c = u_c \cdot x_c, \quad (10)$$

where u_g, u_s, and u_c are the eigenspaces and x_g, x_s, and x_c are the corresponding vectors. The components of the vectors are gender, speaker, and session factors. All eigenspaces are computed using the simplified EM algorithm [12] in a step-by-step manner and only means are considered. All random vectors (e.g., x_g, x_s, and x_c) for each session in the test phase are estimated using the same algorithm. The likelihood ratio score in [13] is used to make the decision as to accept or reject the hypothesis that the utterance was spoken in a particular dialect.

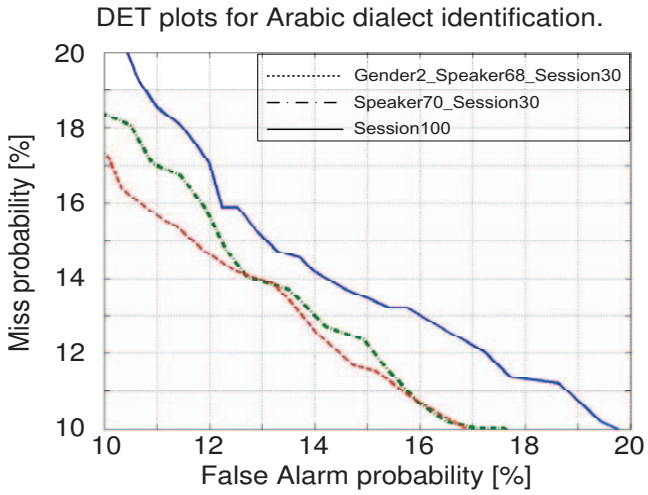

Fig. 1. DET plot of Arabic dialect identification using IIFA models with a total 100 factors.

Table 1. Performance of dialect identification over IIFA models with the same total of the factors.

IIFA model	EER (%)
Session100 (N=2)	14.09
Speaker70_Session30 (N=3)	13.61
Gender2_Speaker68_Session30 (N=4)	13.46

5. EXPERIMENTS

All experiments use the shifted-delta-cepstra (SDC) feature [14]: 7 MFCC coefficients (including coefficient C0) concatenated with SDC 7-1-3-7, which totals 56 coefficients per frame. A UBM with 2048 mixtures was trained via ML criteria, and adapted into five dialect dependent models (UAE, Egypt, Syria, Iraq, and Palestine) using MAP with 20 EM iterations with a relevance factor τ of 14. In Table 1, we show the results for various designs of the IIFA model with the same total 100 factors. The full DET plots for the IIFA models with EERs from the table are shown in Fig 1. In the IIFA model,when $N = 2$ (dialect factor component m_d was obtained using MAP as before), the model reduces to the eigenchannel model with 100 session factors (e.g., "Session100"); when $N = 3$, the model includes 70 speaker and 30 session factors (e.g., "Speaker70_Session30"); when $N = 4$, the model includes 2 gender, 68 speaker, and 30 session factors (e.g., "Gender2_Speaker68_Session30").

It is noted that the IIFA model with speaker and session factor components can produce significantly better results than the model with only the session factor component with the same total factor number. Further gain is obtained with the application of all three factor components (gender, speaker,

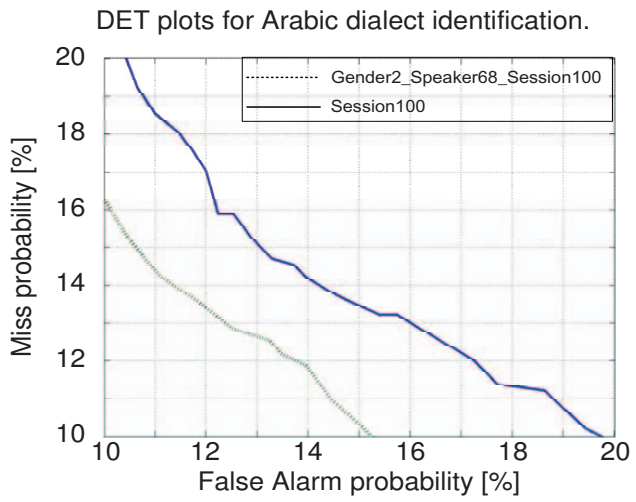

Fig. 2. DET plot of Arabic dialect identification using IIFA models with 100 session factors.

Table 2. Performance of dialect identification over IIFA models with the same number of the session factors.

IIFA model	EER (%)
Session100	14.09
Gender2_Speaker68_Session100	12.77

and session). Since there are only two factors representing gender, the improvement of IIFA with three factor components is not as large, but there is still improvement over the two factor components case.

Here, the number of session factors in the eigenchannel model is 100, while, the number of session factors is only 30 in the IIFA model. So next, we change the number of session factors in the IIFA model from 30 to 100, which is equal to the number of session factors in the eigenchannel model. Fig.2 shows DET performance of eigenchannel and IIFA with $N = 4$ when the number of session factors is the same in both cases. In the figure, the number of the session factors is kept at 100 in the eigenchannel model, while the dimension of the session factors is set to 100 instead of 30 as in the IIFA model (gender and speaker factors are kept), which makes the same number of session factors estimated on the utterance for both models (e.g., "Gender2_Speaker68_Session100"). Significant improvement is obtained with the IIFA model versus eigenchannel as seen in Table2 (i.e., 14.09% vs. 12.77% EER).

6. CONCLUSION

This study has shown that IIFA modeling provides an effective mismatch compensation for GMM-based dialect identification. The major advantage of IIFA modeling is more effective utilization of the transcript/stream knowledge. It should be emphasized that only general stream knowledge is employed, not the specific word or phoneme information for either train or test. The approach suppresses negative components and keeps positive components with a more clear description of the composition of the utterance. Application of IIFA modeling in Arabic dialect identification obtains a 4.23% relative improvement versus eigenchannel with the same total number of factors; and 9.37% relative improvement versus eigenchannel with the same number of session factors. Future research will consider applying the proposed modeling approach for speaker recognition to characterize components of the speaker while employing less factors.

7. REFERENCES

[1] P.A. Torres-Carrasquillo, T.P. Gleason, and D.A. Reynolds, "Dialect identification using Gaussian mixture models," Proc. Odyssey: Speaker & Lang. Recog. Work., 2004.

[2] R Huang and J.H.L. Hansen, "Unsupervised discriminative training with application to dialect classification," *IEEE Trans. SAP*, vol. 15, pp. 2444–2453, 2007.

[3] Y. Lei and J. Hansen, "Dialect classification via discriminative training," INTERSPEECH, 2008, vol. 1, pp. 37–40.

[4] G. Choueitor, G. Zweig, and P. Nguyen, "An empirical study of automatic accent classification," ICASSP, 2008.

[5] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. SAP*, vol. 13, pp. 345–354, May 2005.

[6] P. Kenny P. Oueleet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. ASLP*, vol. 16, pp. 980–988, July 2008.

[7] F. Castaldo, E. Dalmasso, P. Laface, D. Colibro, and C. Vair, "Language identification using acoustic models and speaker compensated cepstral-time matrices," ICASSP, 2007.

[8] P. Torres-Carrasquillo, D. Sturim, D. Reynolds, and A. McCree, "Eigen-channel compensation and discriminatively trained Gaussian mixture models for dialect and accent recognition," INTERSPEECH, 2008.

[9] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," ICASSP, 2004.

[10] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition," Odyssey Workshop, 2006.

[11] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernochy, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. SAP*, vol. 15, pp. 1979–1986, Sep. 2007.

[12] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," INTERSPEECH, 2005.

[13] W. Shen and D. Reynolds, "Improved gmm-based language recognition using constrained MLLR transforms," ICASSP, 2008.

[14] P. Pavel, L. Burget, P. Schwarz, and J. Cernock'y, "Brno University of Technology system for NIST 2005 language recognition evaluation," Proc. of Odyssey, 2006.