Inter-task System Fusion for Speaker Recognition

M. Ferras, S. Madikeri, S. Dey, P. Motlicek and H. Bourlard

Idiap Research Institute, CH-1920 Martigny, Switzerland

{marc.ferras,srikanth.madikeri,subhadeep.dey,petr.motlicek,herve.bourlard}@idiap.ch

Abstract

Fusion is a common approach to improving the performance of speaker recognition systems. Multiple systems using different data, features or algorithms tend to bring complementary contributions to the final decisions being made. It is known that factors such as native language or accent contribute to speaker identity. In this paper, we explore inter-task fusion approaches to incorporating side information from accent and language identification systems to improve the performance of a speaker verification system. We explore both score level and model level approaches, linear logistic regression and linear discriminant analysis respectively, reporting significant gains on accented and multi-lingual data sets of the NIST Speaker Recognition Evaluation 2008 data. Equal error rate and expected rank metrics are reported for speaker verification and speaker identification tasks.

Index Terms: speaker recognition, inter-task, fusion, speaker, accent, language

1. Introduction

In the last decade, speaker recognition system have progressively improved their performance under more and more challenging scenarios. A large number of approaches to speaker recognition have been proposed such as MAP-adapted Gaussian Mixture Models (GMM) [1], Gaussian mean supervectors [2], maximum-likelihood linear regression coefficients [3], Joint Factor Analysis [4], i-vectors [5] with Probabilistic Linear Discriminant Analysis [6], obtaining lower and lower error rates in more and more challenging conditions. Still, system fusion remains a major source of performance improvement that consistently provides gains over its individual subsystems. Indeed, fusion has become a widespread practice due to its simplicity and large improvements obtained, as reported in NIST Speaker Recognition Evaluation (SRE) campaigns [7, 8].

Fusion of speaker recognition systems has been approached using diverse techniques, basically combining features, models or scores. Fusing scores is now ubiquitous for the simplicity, speed and performance improvements obtained. Scores from multiple speaker recognition systems have been fused using weighted average [9, 10] with logistic linear regression having been used for obtaining calibrated likelihood ratios while estimating the optimal fusion weights [11]. Fusion addressing high and low level information have also been explored, with most relevant work integrating high-level and low-level speaker information into a speaker recognition system [12, 13, 14]. All these approaches combine systems aimed at modeling and recognizing speakers, what we name intra-task system fusion. On the other side, inter-task fusion for speaker recognition, that is fusing systems performing speaker recognition together with systems addressing other tasks such as native accent recognition or language recognition has been barely addressed in the past. The most relevant work on including side information into a speaker recognition system can be found in [15], where a neural network is used to combine speaker recognition scores and duration, channel type and SNR features.

The main challenge in combining speaker and side information is that the latter does not provide information about the speaker in a direct usable form. On the other side, it is well known that characteristics such as native language are related to speaker identity. In this paper, we explore two approaches to inter-task fusion for speaker recognition tasks. Score-level and a model-level techniques combining speaker information together with accent and language information are proposed. The rationale behind this form of fusion is that accent and language information can effectively bias the evidence of speech data being uttered by a given speaker.

The paper is organized as follows. Section 2 motivates and discusses the proposed fusion schemes. In Section 3, the speaker verification and accent and language identification systems are described. Section 4 describes the experiments and gives results for the presented fusion methods. Section 5 gives some conclusions on this study.

2. Inter-task Fusion

Speaker recognition systems typically resort to fusing multiple systems to further improve their performance. While it is common to fuse systems developed for speaker recognition tasks, fusion with systems developed for other tasks has not yet been explored. In this paper, we define *inter-task fusion* as the fusion of systems targetting the recognition of speakers and systems targetting the recognition of other modalities such as accent and language.

While spoken accent and language are intuitively related to speaker identity, e.g. native language dominating the accent in other spoken languages, how these should interact with genuine speaker information is not straightforward. Without any other cue, it seems just wrong, for instance, to increase the evidence of a speaker speaking based on non-speaker cues.

In this paper, we explore two distinct approaches to inter-task fusion, namely score fusion and model level fusion. These are described in the following two sections.

This work was supported by the European Union under the FP7 Integrated Project SIIP (Speaker Identification Integrated Project). The authors gratefully thank the EU for their financial support and all project partners for a fruitful collaboration.

2.1. Score-level Fusion

Score level fusion is grounded on fusion techniques developed for speaker verification such as linear logistic regression (LLR) [11]. In this framework, fused scores s_f are obtained for N systems as

$$s_f = b + \sum_{i=1}^{N} w_i s_i, \tag{1}$$

where positive scores favor the same-speaker hypothesis, H_{tar} and negative scores favor the different-speaker hypothesis H_{non}. Besides LLR increasing the speaker discrimination of the fused scores, the technique also calibrates such scores such that

$$s_f = \log \frac{p(s_f|H_{tar})}{p(s_f|H_{non})}. \tag{2}$$

In the proposed fusion approach, rather than scores from several speaker verification systems being fused, side information scores are used from accent or language identification systems. However, there is no reason to believe that the direct linear combination of scores from non-speaker modalities, say accent identification scores, can bias speaker scores in a meaningful way for speaker verification. Indeed, these are meaningful for a non-speaker oriented task and they should ideally behave as noise for a speaker recognition task.

In this work, we assume that speaker verification scores are the main source for making speaker verification decisions, while accent or language identification scores act as side information scores. These are assumed to be available for all speech segments in a speech database. Prior accent probabilities for each enrolled/target speaker are assumed to be known, e.g. estimated from multiple enrolment utterances from ground truth data. Accent posterior probabilities are estimated using an accent identification system for the test segments. This fits a speaker identification scenario where a test utterance is to be compared to many speakers in a database that has been enriched with accent metadata.

While increasing speaker evidence based on non-speaker information seems like a dangerous practice, the opposite appears as intuitive and meaningful. The rationale is simple: non-matching enrolment and test accents are indicative of a non-target speaker verification score. We understand this to be a hard-decision filtering approach to inter-task fusion that is straightforward to implement by setting an extreme negative fused score in case of accent mismatch. In practice, accent and language identification systems make errors that propagate into the fused system decisions. An approach performing soft-filtering, e.g. slightly decreasing speaker verification scores based on accent/language mismatch, may be better suited for a real-world application. The latter is adopted in this work.

Fig. 1 shows two block diagrams for soft-filtering score fusion using two types of side information scores, focusing on the accent modality. Fig. 1(a) uses binary accent mismatch values as scores while Fig. 1(b) uses accent mismatch values in the range [0,1], namely the product of enrollment and test accent probabilities as scores. The soft filtering is performed by optimizing the fusion weights that minimize Cllr cost function [16] while calibrating the scores. In both schemes, side information scores are zero when enroll and test accents are the same. This prevents the fusion system from increasing the speaker verification scores based on non-speaker information. Any performance improvement arises from scores for which accents are mismatched.

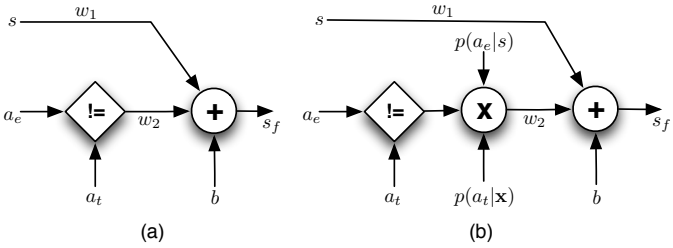

Figure 1: Logistic Linear Regression (LLR) binary and probabilistic approaches to fusion of speaker verification scores and accent side information. a_e and a_t are the known most probable accents for the enrolled speaker and test utterance respectively: (a) Fusion with a binary indicator of accent mismatch. (b) Given accent mismatch between enrolment and test, fusion with the product of accent probability for the enrolled speaker, $p(a_e|s)$, times the posterior probability for the test, $p(a_t|\mathbf{x})$.

The accent identification system is assumed to output a likelihood score $p(\mathbf{x}|a)$ for each accent a being considered. Such scores are transformed into posterior probabilities as $p(a|\mathbf{x}) = \frac{p(\mathbf{x}|a)p(a)}{\sum_{a'} p(\mathbf{x}|a')p(a')}$. The maximum a posteriori accent for a given test segment t is taken for fusion according to $a_t = \arg\max_a p(a|\mathbf{x})$. For an enrolled speaker s the most probable accent is taken according to $a_e = \arg\max_a p(a|s)$

An analogous approach to speaker-accent score level fusion is followed for language side information, just by replacing accent by language, as $p(\mathbf{x}|a)$ by $p(\mathbf{x}|l)$ and $p(a|\mathbf{x})$ by $p(l|\mathbf{x})$, where l is any of the languages considered in the experiments.

2.2. Model-level Fusion

In this work, i-vector based systems are used for speaker verification, accent identification and language identification systems. These systems purposely share the same spectral envelope features, the same Universal Background Model (UBM) and T matrix for i-vector extraction. I-vectors are post-processed using Linear Discriminant Analysis (LDA), length normalization and finally scored using Probabilistic Linear Discriminant Analysis (PLDA). This architecture is shown in Fig. 3. Systems developed for different tasks only differ in the LDA and PLDA back-end, which are trained using the speaker, accent and language labels for speaker verification, accent identification and language identification tasks, respectively. The dimensions of the projected i-vectors may differ as well, especially for LDA.

The LDA and PLDA frameworks are able to project the i-vector space onto speaker, accent and language discriminative subspaces while still resulting in high performance systems. In this work, we propose to use the LDA framework to perform i-vector fusion as well. To fuse the speaker verification and accent identification systems, i-vectors are projected with LDA models trained specifically for speaker and accent discrimination. The projected vectors are concatenated and used to train a conventional speaker verification back-end with LDA followed by PLDA with only speaker labels (that is, accent labels is discarded). The concatenation of heterogeneous i-vectors complements the speaker models with accent specific information. This scheme is shown in Fig. 2. The same procedure is applied

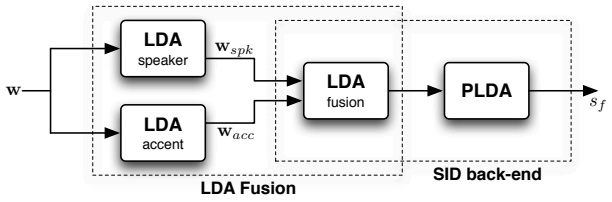

Figure 2: Fusion of speaker i-vectors $\mathbf{w}_{spk}$ and accent i-vectors $\mathbf{w}_{acc}$ via LDA prior to PLDA scoring.

to fuse language information into speaker verification systems. In this case, the accent LDA model is simply replaced by the language LDA model.

Such an approach to fusion has several advantages. First, it does not depend on hard decisions being made, thus preventing error propagation to the fused scores. Speaker, accent and language information is represented as the continuous components of i-vectors, where fusion is performed. A second advantage for LDA fusion is that it is fully data driven. In principle, this compares favorably with the score fusion approach relying on matching side information such as accent or language. The latter, although reasonable, are indeed inflexible assumptions.

3. Experimental Setup

To evaluate the approaches to inter-task fusion described above for accent and language side information, several systems are required: a speaker verification system, an accent identification system and a language identification system. These three systems have been developed using the same i-vector framework, shown in Fig. 3, using LDA for post-processing and PLDA for scoring. The LDA and PLDA modules are trained using speaker, accent or language labels respectively to target the different tasks.

Feature vectors consist of 19 Mel-Frequency Cepstral Coefficients extracted every 10ms over a 25ms sliding window and post-processed using feature warping over a 3s long sliding window. A 2048 Gaussian component UBM was trained on the Fisher female data using maximum-likelihood estimation. The same data was used to train the $\mathbf{T}$ matrix for i-vector estimation, extracting i-vectors of 400 dimensions.

For the speaker verification LDA+PLDA back-end, NIST SRE'04, SRE'05, SRE'06 and Fisher female data were used for training. For the accent identification back-end, only utterances in English language with native speakers of English, Chinese, Hindi, Russian and Korean were used from the same data set. For the language back-end, only utterances in English, Chinese, Hindi, Russian and Korean languages were used from the same data set plus NIST LRE'05 and LRE'07, plus the Callfriend data set. Accents and languages were unevenly represented in the training data with English and Chinese being over represented when compared to scarce data for Hindi and Korean. Prior to fusion, the speaker i-vector dimension was unchanged and while the accent and language i-vector dimensions were reduced from 400 down to 4. To train the LDA after concatenating speaker i-vectors and accent/language i-vectors as the case may be, we discarded data for which there were no speaker labels.

The performance of these systems was evaluated on the NIST SRE'08 condition 6 data, consisiting of multi-language data for the 5 languages above, and condition 7 data, consisting

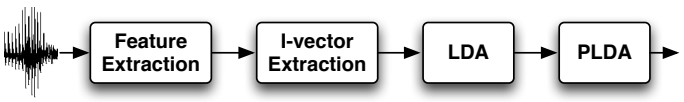

Figure 3: I-vector system using PLDA scoring. The LDA and PLDA modules are trained to discriminate labels defined by the task, i.e. speaker, accent or language labels.

of accented English data spoken by native speakers of the same 5 languages. To evaluate the performance of the fusion methods of Section 2 on both speaker verification and speaker identification tasks, accented and multi-language trial lists were generated by scoring each target speaker against 1 target and 300 randomly chosen non-target utterances, for a total of 150'000 and 250'000 trials used for accented and multi-language evaluation respectively.

Table 1 shows the EER of the baseline verification system evaluated on accented English (c7) and multi-language speech data (c6). Accent and language identification accuracies are also given for the respective data sets. From these figures, it seems that accent identification is a tougher task compared to language recognition. This makes sense intuitively as accent identification deals with intra-language variation whereas language identification deals with inter-language variation. However, it must be emphasized that accented English data was rather scarce for native speakers of Hindi and Korean languages.

System	**SRE'08 c7**	**SRE'08 c6**
Speaker Verification	1.87% EER	1.34% EER
Accent Identification	87.36% Acc.	-
Language Identificaiton	-	94.63% Acc.

Table 1: Baseline speaker verification (in EER), and accent and language identification (identification accuracy) performances.

For LLR fusion, the Bosaris toolkit [17] was used to optimize the fusion parameters. The Kaldi toolkit [18] was used for speaker verification, accent identification and language identification system development, including LDA fusion and PLDA scoring.

4. Experiments

A preliminary set of experiments were run to assess the speaker discriminative power of accent and language i-vectors, obtained after projection onto the PLDA inter-accent and inter-language matrices. Table 2 shows EER for systems using accent and language i-vectors on the baseine speaker verification system. EER for accent i-vectors were 14 times worse than for speaker i-vectors, 25 times worse for language i-vectors. These numbers suggest an inverse relation w.r.t. the accent and language identification accuracy of systems using such i-vectors. The discrimination amongst speakers is deemed residual in both cases.

A series of experiments were run to assess different methods of including accent information in a speaker verification system on the accented English data of the SRE'08 condition 7. The Equal Error Rate (EER) of the systems are compared for speaker verification tasks. To compare speaker identification systems we define a metric called the Expected Rank (ER), given by $E[r] = \sum_r p(r)r$, where r is the number of speakers

Table 2: Evaluation of accent and language i-vectors on accented (c7) and multi-language (c6) speaker verification tasks.

System	SRE'08 Cond.	EER (%)
Speaker i-vectors	c7	1.87
Accent i-vectors	c7	26.70
Speaker i-vectors	c6	1.34
Language i-vectors	c6	33.67

retrieved and $p(r)$ denotes the probability that the target speaker appears in the top r matches among the enrolled speakers for a test audio. The ER of a system indicates the average rank of the target speaker over a large number of test cases when the system returns the top R speakers closest to the test audio. This metric is defined based on the Mean of Average Precision (MAP) often used in Information Retrieval (IR) systems [19]. Note that ER values are non-negative and can also be less than 1 based on the probability distribution of r.

Table 3 shows results for logistic linear regression (LLR) and LDA fusion methods as described in Section 2 for speaker and accent fusion. For the baseline non-fused system, an absolute EER of 1.87% and an ER of 1.96 were obtained, meaning target speakers are expected to appear as second match in the identification task.

An oracle experiment was run to assess the potential of a filtering approach to fusion. In the Filter Oracle system, speaker verification scores were given an extreme negative score if the ground truth target and test accents were different. As shown in Table 3, 5% (1.87% to 1.77%) and 18% (1.96 to 1.60) maximum relative improvements in EER and ER can be obtained using such filtering approach. Assuming no accent ground truth is available for test utterances, we use the developed accent identification system, obtaining 87.36% identification accuracy, to estimate accents. When the maximum-a-posteriori accent is used for each test utterance, EER can decrease from 1.87% to 1.80% with LLR based score fusion, relatively close to the oracle EER of 1.77%. Both LLR Binary and LLR Prob. approaches, using either binary values or the product of target and test accent probabilities as fusion scores, yield the same results. Regarding ER, LLR fusion achives 7% relative improvement (1.97 to 1.83), but remains far from 1.60, obtained using the ground truth accent information.

LDA fusion, shown in the last row of Table 3, achieves 1.80% EER, the same as the score fusion methods above. ER is the smallest of all methods, obtaining a 37% relative decrease from the baseline, from 1.97 to 1.23.

According to these experiments, significant improvements can be achieved using accent information for speaker verification, and especially for a speaker identification task using LDA fusion, that does not make any assumptions or require any decisions to be made. Both methods reach a minimum EER which is close to that obtained using the ground truth accent information. This suggests an upper bound on the EER improvements obtained by using accent information, although more effective fusion methods may overcome the limits observed in this work.

A second set of experiments were run for language and speaker fusion on the multi-language trials of the SRE'08 condition 6. From Table 4, the absolute EER for the baseline is 1.34% considerably lower than the 1.87% EER obtained for the multi-accent trials of condition 7.

Table 3: Fusion of a speaker verification system with an accent identification system. Equal Error Rate and expected rank for speaker verification and identification tasks are shown for several fusion approaches.

System	EER (%)	ER
No fusion	1.87	1.97
Filter Oracle	1.77	1.60
LLR Binary	1.80	1.83
LLR Prob.	1.80	1.81
LDA	1.80	1.23

Regarding score fusion, the Filter Oracle experiment for the language factor shows minimal EER improvements over the baseline. We believe this is due to the fact that language is not an identity determining factor as accent may be. Indeed, many people can speak multiple languages, and not speaking the enrolled languages is less constraining than not speaking a certain accent during enrollment, which translates in fusion affecting very few scores . Being inspired in filtering, LLR fusion shows similar trends, with minor improvements in both EER and ER and even some loss for the LLR Prob. system.

LDA fusing speaker and language i-vectors obtains gains over 8% EER in relative terms, from 1.34% to 1.23%, over the baseline. In this case, a fully data driven approach to fusion not relying on language identification hard decisions is definitely advantageous. Minor improvements in ER are observed as well.

Table 4: Fusion of a speaker verification system with a language identification system. Equal Error Rate and expected rank for speaker verification and identification tasks are shown for several fusion approaches.

System	EER (%)	ER
No fusion	1.34	0.64
Filter Oracle	1.33	1.64
LLR Binary	1.33	0.65
LLR Prob.	1.39	0.66
LDA	1.23	0.63

5. Conclusion

In this paper we explored several approaches to score level and model level fusion of speaker and non-speaker information, namely accent and language, to improve speaker verification and identification performance. Both score and model level approaches provided improvements in both equal error rate and expected rank. Score level fusion, being based on filtering and relying on accent and language identification decisions, obtained small improvements for language fusion. Since a speaker can speak several languages, language mismatch is not a decisive factor to decrease the speaker scores, at least not as decisive as accent. The proposed i-vector fusion based on a data-driven approach such as linear discriminant analysis outperformed score level fusion in most cases, and especially for language-speaker fusion.

6. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[2] W. M. Campbell, D.E. Sturim, and D. A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[3] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR Transforms as Features in Speaker Recognition," in *Proc. EUROSPEECH*, September 2005, pp. 2425–2428.

[4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2008.

[5] N. Dehak, P. Kenny, R. Dehak, O. Glembek, P. Dumouchel, L. Burget, V. Hubeika, and F. Castaldo, "Support vector machines and joint factor analysis for speaker verification," in *Proc. IEEE ICASSP*, Taipei, Taiwan, 2009.

[6] Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Proc. INTERSPEECH*, 2011, pp. 249–252.

[7] "Nist speaker recognition evaluation 2010," `http://www.nist.gov/itl/iad/mig/sre10.cfm`, Accessed March 25th, 2016.

[8] "Nist speaker recognition evaluation 2012," `http://www.nist.gov/itl/iad/mig/sre12.cfm`, Accessed March 25th, 2016.

[9] Kevin R Farrell, Ravi P Ramachandran, and Richard J Mammone, "An analysis of data fusion methods for speaker verification," in *Proc. IEEE ICASSP*, 1998, vol. 2, pp. 1129–1132.

[10] Jon Atli Benediktsson and Philip H Swain, "Consensus theoretic classification methods," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 22, no. 4, pp. 688–704, 1992.

[11] Niko Brümmer, Lukáš Burget, Jan Honza Černockỳ, Ondřej Glembek, František Grezl, Martin Karafiat, David A Van Leeuwen, Pavel Matě, Petr Schwarz, and Albert Strasheim, "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2072–2084, 2007.

[12] Douglas Reynolds, Walter Andrews, Joseph Campbell, Jiri Navratil, Barbara Peskin, Andrea Adami, Qin Jin, Dalibor Klusacek, Joy Abramson, Radu Mihaescu, et al., "The supersid project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. IEEE ICASSP*, 2003, vol. 4, pp. IV–784.

[13] Sachin S Kajarekar, Nicolas Scheffer, Martin Graciarena, Elizabeth Shriberg, Andreas Stolcke, Luciana Ferrer, and Tobias Bocklet, "The sri nist 2008 speaker recognition evaluation system," in *Proc. IEEE ICASSP*, 2009, pp. 4205–4208.

[14] K Murty and Bayya Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *Signal Processing Letters, IEEE*, vol. 13, no. 1, pp. 52–55, 2006.

[15] William M Campbell, Douglas A Reynolds, Joseph P Campbell, and Kevin Brady, "Estimating and evaluating confidence for forensic speaker recognition.," in *Proc. IEEE ICASSP*, 2005, pp. 717–720.

[16] Niko Brümmer and Johan Du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2, pp. 230–275, 2006.

[17] "Bosaris toolkit," `https://sites.google.com/site/bosaristoolkit/`, Accessed March 25th, 2016.

[18] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," 2011.

[19] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al., *Introduction to information retrieval*, vol. 1, Cambridge university press Cambridge, 2008.