# PARALLEL TRANSFORMATION NETWORK FEATURES FOR SPEAKER RECOGNITION

Alberto Abad[1], Jordi Luque[2], Isabel Trancoso[1,3]

[1] $L^2F$ - Spoken Language Systems Laboratory, INESC-ID Lisboa, Portugal
[2] TALP research center, Universitat Politècnica de Catalunya, Spain
[3] Instituto Superior Técnico, Lisboa, Portugal
alberto.abad@inesc-id.pt

## ABSTRACT

The use of speaker adaptation transforms as features for speaker recognition is an appealing alternative to conventional short-term cepstral features. In general, this kind of methods are language dependent and limited by the need of speech recognition in the client speakers language. In this paper, we generalize a recently proposed method –named Transformation Network features with SVM modeling– in order to become language independent and overcome the need for accurate speech recognition. This is accomplished by using a set of parallel acoustic models in several different languages to obtain a high-dimensional Parallel Transformation Network feature vector for speaker characterization.

***Index Terms***— Speaker recognition, transformation features, connectionist adaptation

## 1. INTRODUCTION

Most successful current state of the art speaker identification systems are commonly based on the combination of several different methods [1, 2, 3]. Particularly, many efforts have been recently devoted to investigate new alternatives to conventional short-term cepstral based methods. One of the main motivations is the need for dealing with the inability of short-term features – extracted from few milliseconds – for capturing higher order structure information in speech that might be useful for characterizing speakers.

In [4] an appealing method that uses Maximum-Likelihood Linear Regression (MLLR) speaker adaptation transform based features for speaker modeling is proposed. Instead of modeling cepstral observations directly, it models the "difference" between the speaker-dependent and the speaker-independent models. The high-dimensional vectors formed by the transform coefficients are then modeled as speaker features using support vector machines (SVM). More recently, in [5], we have proposed a solution to make use of speaker adaptation derived features of a connectionist hybrid artificial neural network/hidden Markov model (ANN/HMM) [6] speech recognizer. Our approach uses a method known as Transformation Network (TN) [7] to train a linear input network that maps the speaker-dependent input vectors to the speaker independent system, while keeping all the other parameters of the neural network fixed.

In general, approaches based on speaker adaptation features like [4] and [5] present some inherent limitations. On the one hand, they are language dependent, since speech recognition is necessary for speaker transformation estimation. On the other hand, as we showed in [8], they are very sensitive to the quality of the automatic speech transcriptions used for speaker adaptation.

In the present work, we go a step forward and generalize the recently proposed Transformation Network features with SVM modeling (TN-SVM) approach in order to be language independent and without the need for accurate transcriptions. Somehow similarly to the well-known Phone Recognition followed by Language modeling (PRLM) [9] approach for Language Identification, a set of independent speech recognition networks of different arbitrary languages are used to phonetically decode the speaker data and to estimate the speaker TN adapted to every language network. Then, every language-dependent TN feature vector is composed in a larger Parallel Transformation Network (PTN) vector, which is used for speaker recognition.

The paper is organized as follows. In the next section the TN-SVM approach for speaker recognition is briefly described. Section 3 reviews the main limitations of the TN-SVM technique and a possible solution based on the use of parallel speaker acoustic transformations is proposed. The experimental assessment of the PTN-SVM novel approach and its comparison to Gaussian Supervector (GSV) and to TN-SVM baseline systems for the English and non-English trials of a sub-set of one NIST Speaker Recognition Evaluation 2008 [10] test condition is reported in Section 4. Finally, we present the conclusions in Section 5.

## 2. TN FEATURES FOR SPEAKER RECOGNITION

### 2.1. ANN/HMM speech recognition and phone transcriptions

Our core speech recognizer uses Multiple Layer Perceptron (MLP) networks that act as phoneme classifiers for estimating the posterior probabilities of a single state Markov chain monophone model. The baseline system for narrowband data combines four MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), log-RelAtive SpecTrAl features (RASTA, 13 static + first derivative), Modulation SpectroGram features (MSG, 28 static) and the advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives). The number of context input frames is 13 for the PLP, RASTA and ETSI networks and 15 for the MSG network. The decoder of the recognizer is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition.

The method followed for generating phoneme transcriptions of the speaker data is crucial for speaker recognition performance of the proposed method. This procedure is even more important when a speech recognition system that is not well-matched to the task char-

This work was partially supported by FCT (INESC-ID multi-annual funding) through the PIDDAC Program funds, and also through project CMU-PT/HuMach/0039/2008.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

acterisitics is used. In our previous work, we have used MLP acoustic models trained with down-sampled American English Broadcast News (BN) data. A possibility to generate phonetic transcriptions consists of forced alignment with high-quality word-level automatic transcriptions provided by NIST [5]. Alternatively, the same acoustic models can be used with an *n-gram* phone model to obtain phone transcriptions [8].

### 2.2. MLP/HMM Speaker Adaptation

The Transformation Network [7] technique employs a trainable linear input network to map the speaker dependent (SD) input vectors to the characteristics of the speaker independent (SI) connectionist system. In order to train the TN for a new speaker, the weights of the mapping are initialized to an identity matrix. This guarantees that the SI model is the initial point prior to adaptation. During training, the output error of the posterior probabilities is calculated and back-propagated as usual in MLP training. But the SI part is kept *frozen* and weight adaptation is performed only in the new transformation network. The result is a linear mapping that represents the differences between a new speaker and a generic SI model. Notice, that although it can be considered a sort of spectral normalization technique it presents some particularities. First, the TN method does not impose any restriction at the LIN output in terms of a reference or target speaker. The only restriction comes from the output error minimization. Second, the TN approach is architecture dependent, hence it can not be considered a generalized spectral normalization technique.

### 2.3. TN feature extraction and SVM speaker modeling

Linear speaker dependent mappings for each data segment are independently trained and consequently a speaker adapted transformation matrix is obtained for each segment. In order to avoid capturing too much information of the background or channel conditions, long segments of silence were removed from the adaptation data. All the data of the speech segment is used for estimating the transformation matrix (no data is kept for cross-validation). A fixed number of training epochs with a relatively small adaptation step is used for estimating the transformation weights. In this work we estimate tied networks sharing the same weights for all frames independently of their position in the input context instead of training a full-matrix. Thus, the dimensionality of the linear mapping is reduced to just $[N_{feat}, N_{feat}]$. The coefficients from the linear mapping obtained for each speaker are concatenated in a vector together with the segment mean and variance statistics of the feature data. The complete TN feature vector is obtained as the concatenation of each individual network vector (PLP, RASTA, MSG and ETSI) resulting in a vector of size 3895.

The connectionist transformation network feature vectors are used to train SVM target speaker models. Linear kernel is used for training speaker models and min-max normalization in the [0,1] rank is applied.

## 3. GENERALIZED PARALLEL TN APPROACH TO SPEAKER RECOGNITION

### 3.1. TN-SVM limitations

The TN-SVM approach described in the previous section presents some limitations that are generally common to the approaches that use speaker adaptation features for speaker recognition. The most important one is that the proposed system is language-dependent: a well-performing ASR system of the language used by the client speakers is needed for both phonetic transcription and for acoustic model adaptation.

A partial solution to the need of manual –or at least high quality– transcriptions when a high performance ASR system is not available was proposed consisting of the use of a phone *n-gram* model (or a phone-loop grammar) to obtain a coarse phonetic transcription that can be later used for acoustic model adaptation [8]. Although some speaker information was retained, a considerable performance drop results from this approach. This performance loss is partially due to the lack of high-performing acoustic models for the task (conversational telephone speech), but also to the presence of speech of other unknown languages different from American English.

### 3.2. Towards language-independence: Parallel Networks

The new approach proposed in this work is partially inspired by the Parallel Phone Recognition followed by Language Modeling (Parallel PRLM) approach for language recognition [9]. The PRLM technique uses a phone tokenizer of any arbitrary language to extract the phonotactics (relationship between phones) of every target language in the phonetic space of that phonetic classifier. Using several parallel tokenizers of different languages permits extracting the phonotactics mapped to different language phonetic spaces and their combination provides significant language recognition improvements.

Similarly, in this work we explore the use of several acoustic models of different languages in parallel to obtain individual language-dependent speaker transformations. The rationale is that the acoustic characteristics of each individual speaker can be mapped to the phonetic acoustic space of any arbitrary language and that this mapping may retain speaker-dependent information. The total vector formed by the speaker transformations in every different language is called Parallel Transformation Network vector. As a result of the combination of the different acoustic model adaptation parameters, the PTN vector is expected to provide speaker discriminant information that is language independent and it is also expected to significantly improve the phone-loop TN-SVM method that uses a single acoustic model for American English.

#### *3.2.1. Acoustic MLP networks description*

Four acoustic model language-dependent MLP networks have been considered: the American English (*en*) models of the previous section 2 and acoustic models for European Portuguese (*pt*), Brazilian Portuguese (*br*) and European Spanish (*es*). The *en* system was trained with the HUB-4 96 and HUB-4 97 down-sampled data sets, that contain around 142 hours of TV and Radio Broadcast data. The *pt* phonetic classifier was trained with 57 hours of BN down-sampled data, and 58 hours of mixed fixed-telephone and mobile-telephone data. The *br* models were trained with around 13 hours of BN down-sampled data. The *es* networks used 36 hours of BN down-sampled data and 21 hours of fixed-telephone data. In this work, only monophone units are modeled, resulting in MLP networks of 41 (39 phonemes +1 silence + 1 respiration) soft-max outputs in the case of *en*, 39 for *pt* (38 phonemes + 1 silence), 40 for *br* (39 phonemes + 1 silence) and 32 for *es* (31 phonemes + 1 silence). All the neural networks are composed by two hidden layers of 500 units each one. In contrast to our previous work, we have used reduced dimensionality networks due to the fact that we consistently observed improvements of the TN-SVM method when smaller networks were used. The four feature streams (PLP, RASTA, MSG and ETSI) are used by all the language-dependent systems.

### 3.2.2. *PTN feature extraction*

The process for extracting the PTN feature vector is identical to the one described in section 2.3 and it is followed independently for each individual acoustic model network. In order to obtain the phonetic transcription for each language-dependent network, a simple phone-loop grammar of all the phones of the given language is used with phoneme minimum duration of two frames. The use of a *n-gram* phone model was dismissed since it would impose phonetic relations that are characteristic of the network language. Once the phonetic transcription is obtained, the speaker transformation is estimated, and the TN feature vector obtained for each language network. The final PTN vector is of size 15580 (4*3895). These high-dimensionality vectors are used for training SVM speaker models and for testing as described in section 2.3.

## 4. SPEAKER RECOGNITION EXPERIMENTS

### 4.1. Experimental set-up

### 4.1.1. *Task definition, data and performance metrics*

Speaker verification is assessed in one sub-set of the *short2-short3* NIST Speaker Recognition Evaluation 2008 test condition [10]. Concretely, we consider the *telephone-telephone* training and test condition.

The training and testing data sets of the *telephone-telephone* condition consist of two-channel telephone conversational excerpts, of approximately five minutes total duration, with the target speaker channel designated. The gender of speakers in train and test segments is also known. The complete test condition consists of 37050 trials with 648 male and 1140 female target speakers, each of them being tested against approximately 20 different test segments. Test trials are classified in English trials (18360) and non-English trials (18690). Additional training data sets from previous SRE evaluations are used for the development of the speaker recognition system. Single channel conversation sides of approximately 5 minutes of SRE2004, SRE2005 and SRE2006 evaluations are used for background modelling/training. 200 speech segments (100 female and 100 male) from SRE2004 and SRE2005 are used for T-norm score normalization.

The detection cost function (DCF) is the metric used in this work with the parameter values of NIST 2008 evaluation campaign, that is, $P_{target}$=0.01, $C_{miss}$=10, $C_{FalseAlarm}$=1. The minimum and actual normalized DCF point (min/act $C_{Norm}$) are provided for assessment of the speaker detection systems. Additionally, we also report the Equal Error Rate (EER) and the Detection Error Tradeoff (DET) curve for a better evaluation of the speaker recognition systems under study.

### 4.1.2. *Score calibration*

Every single system is calibrated with the *s-cal* tool available in the Focal toolkit [11]. It allows to discriminatively train a mapping to convert detection scores to detection log-likelihood-ratios. Linear logistic regression is further applied to the s-calibrated scores. All calibration parameters are gender-dependent. A five-fold cross-validation strategy on the test set is applied to simultaneously estimate the calibration parameters and to evaluate speaker detection systems.

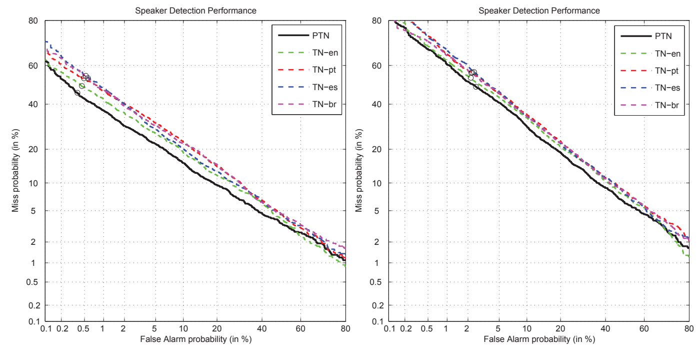

**Fig. 1**. *DET curves of the language-dependent TN systems (TN-en, TN-pt, TN-es and TN-br) and of the proposed PTN system for only English (left) and only non-English (right) trials.*

| | English | | non-English | |
|---|---|---|---|---|
| System | min/act $C_{Norm}$ | EER | min/act $C_{Norm}$ | EER |
| PTN | 0.464/0.496 | 12.53 | 0.699/0.744 | 18.67 |
| TN-en | 0.517/0.539 | 14.23 | 0.719/0.753 | 20.16 |
| TN-pt | 0.562/0.583 | 16.02 | 0.745/0.791 | 21.18 |
| TN-es | 0.578/0.599 | 14.81 | 0.766/0.806 | 20.92 |
| TN-br | 0.567/0.591 | 15.87 | 0.746/0.782 | 20.89 |

**Table 1**. *Minimum and actual $C_{Norm}$ and EER (%) of the language-dependent TN systems (TN-en, TN-pt, TN-es and TN-br) and of the proposed PTN system for only English and only non-English trials.*

### 4.1.3. *Contrastive systems description*

A Gaussian Supervector (GSV) system and the TN-SVM baseline system that uses the NIST transcriptions for phonetic forced alignment (TN-base) are used for contrastive purposes.

The GSV system is based in the kernel described in [12] and it is used in this work as a reference for language-independence. Its front-end extracts 19 PLP static features with log-RASTA processing and the frame energy together with the first and second derivatives. For each speech segment low-energy and high likely non-speech frames are removed, and mean and variance feature normalization is performed. Gender dependent UBMs of 1024 components are trained with data obtained from SRE2004 and SRE2005 training data sets. UBM means are adapted with 1 MAP iteration with a relevance factor of 16 to obtain the speaker models.

T-norm score normalization is applied to the detection scores produced by all the systems developed in this work.

### 4.2. Experiments with the individual language networks

The aim of this first set of experiments is to determine the speaker recognition ability of each individual language-dependent network on its own. DET curves, min/act $C_{Norm}$ and EER scores are reported in Figure 1 and Table 1 for English and non-English trials. It can be observed that each individual network is able to provide similar speaker recognition performance. The TN-en system performs considerable better than the three other language-dependent detectors in the case of English trials, as one could expect. In the case of non-English trials, the four network-based features are closer, and only a slightly better performance of the *en* based features is ob-

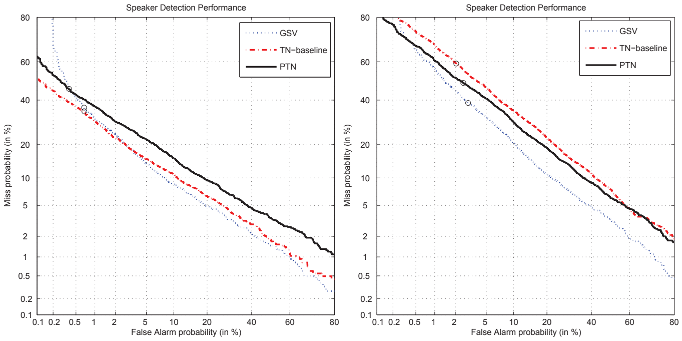

**Fig. 2**. *DET curves of the Gaussian supervector (GSV), the baseline TN-SVM (TN-base) and the proposed PTN systems for only English (left) and only non-English (right) trials.*

| | English | | non-English | |
|---|---|---|---|---|
| System | min/act $C_{Norm}$ | EER | min/act $C_{Norm}$ | EER |
| GSV | 0.406/0.429 | 8.84 | 0.633/0.680 | 14.58 |
| TN-base | 0.394/0.411 | 9.83 | 0.772/0.795 | 21.17 |
| PTN | 0.464/0.496 | 12.53 | 0.698/0.744 | 18.67 |

**Table 2**. *Minimum and actual $C_{Norm}$ and EER(%) of the Gaussian supervector (GSV), the baseline TN-SVM (TN-base) and the proposed PTN systems for only English and only non-English trials.*

served, which is probably due to the fact that is the network trained with more data and that models a larger number of acoustic units. On the other hand, the TN-es system consistently achieves the weakest performance in terms of $C_{Norm}$, which might be due to the reduced number of acoustic classes modeled. In both English and non-English trials cases, the use of the PTN vector combining the transformation parameters of every single language-dependent network provides significant speaker recognition improvements.

### 4.3. Comparison experiments

In these experiments, the GSV system provides a reference for language-independence. Thus, it can be first noticed in the results depicted in Figure 2 and Table 2 that the non-English trials are significantly more difficult, which is in accordance with the evaluation results [10]. On the one hand, it can be observed that the TN-base system is able to provide speaker recognition performance close to the GSV detector in the English trials test set. However, a large performance drop occurs in the non-English trials compared to the GSV. On the other hand, the PTN system provides quite language-independent speaker recognition. Both in the English trials and the non-English trials, the DET curves are at approximately the same log-distance to the GSV curves. In fact, in the case of non-English trials, the new PTN system performs significantly better than the previously proposed TN-SVM baseline.

## 5. CONCLUSIONS

A recenlty proposed approach –named Transformation Network features with SVM modelling– allows the extraction of meaningful features for speaker recognition derived from adaptation techniques used in connectionist ANN/HMM speech recognition. In spite of the previous encouraging results, it has been shown that a considerable performance loss occurs when the method is applied to trials with speech in a language that does not match the one of the speech recognizer. The use of parallel speaker transformation features of different language-dependent acoustic models is proposed to overcome this language dependency limitation and also the need for accurate phonetic transcriptions. Experimental results show that the new Parallel Transformation Network features provide more language independence in the speaker recognition performance, and significantly better results than the previous TN-SVM approach in unmatched language trials.

## 6. REFERENCES

[1] D.E. Sturim, W.M. Campbell, Z N. Karam, D.A. Reynolds, and F.S. Richardson, "The MIT Lincoln Laboratory 2008 Speaker Recognition System," in *Proc. ISCA Interspeech*, 2009, pp. 2359–2362.

[2] S. Kajarekar, N. Scheffer, M. Graciarena, E. Shriberg, A. Stolcke, L. Ferrer, and T. Bocklet, "The SRI NIST 2008 Speaker Recognition Evaluation System," in *Proc. IEEE ICASSP*, 2009, pp. 4205–4208.

[3] L. Burget, M. Fapšo, V. Hubeika, O. Glembek, M. Karafiát, M. Kockmann, P. Matějka, P. Schwarz, and J. Ćernocký, "BUT system for NIST 2008 speaker recognition evaluation," in *Proc. ISCA Interspeech*, 2009, pp. 2335–2338.

[4] A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman, "MLLR transforms as features in speaker recognition," in *Proc. ISCA Interspeech*, 2005, pp. 2425–2428.

[5] A. Abad and J. Luque, "Connectionist Transformation Network Features for Speaker Recognition," in *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2010, pp. 20–27.

[6] N. Morgan and H. Bourlad, "An introduction to hybrid HMM/connectionist continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 25–42, 1995.

[7] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation," in *Proc. ISCA Eurospeech*, 1995, pp. 2183–2186.

[8] A. Abad and I. Trancoso, "Speaker Recognition Experiments using Connectionist Transformation Network Features," in *Proc. ISCA Interspeech*, 2010, pp. 378–381.

[9] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[10] "The NIST year 2008 speaker recognition evaluation plan," http://www.nist.gov/speech/tests/spk/2008/.

[11] N. Brummer, "Focal: Tools for Fusion and Calibration of automatic speaker detection systems," http://www.dsp.sun.ac.za/ nbrummer/focal/.

[12] 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.