Speaker Recognition Experiments using Connectionist Transformation Network Features

Alberto Abad[1] *and Isabel Trancoso*[1,2]

[1] L^2F - Spoken Language Systems Laboratory, INESC-ID Lisboa, Portugal
[2]Instituto Superior Tcnico, Lisboa, Portugal

alberto.abad@inesc-id.pt

Abstract

The use of adaptation transforms common in speech recognition systems as features for speaker recognition is an appealing alternative approach to conventional short-term cepstral modelling of speaker characteristics. Recently, we have shown that it is possible to use transformation weights derived from adaptation techniques applied to the Multi Layer Perceptrons that form a connectionist speech recognizer. The proposed method – named Transformation Network features with SVM modelling (TN-SVM) – showed promising results on a sub-set of NIST SRE 2008 and allowed further improvements when it was combined with baseline systems. In this paper, we summarize the recently proposed TN-SVM approach and present new results. First, we explore two alternative approaches that may be used in the absence of high quality speech transcriptions. Second, we present results of the proposed approach with Nuisance Attribute Projection for session variability compensation.

Index Terms: speaker recognition, transformation features, connectionist adaptation

1. Introduction

Modelling of short-term cepstral features by means of any pattern classification method – typically Gaussian Mixture Models (GMM) [1] or Support Vector Machines (SVM) [2] – is one of the most successful approaches to the speaker recognition task. However, several efforts have been recently devoted to investigate new alternative approaches to conventional short-term cepstral based methods [3]. 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 modelling is proposed. Instead of modelling 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 modelled as speaker features using support vector machines (SVM).

These encouraging results motivated our recent work in [5], where we also make use of adaptation transforms employed in speech recognition as features for speaker recognition. However, in contrast to [4], the automatic speech recognizer that we rely on for computing the “differences” between the speaker independent and the speaker dependent model is a connectionist hybrid artificial neural network/hidden Markov model (ANN/HMM) system [6]. Unfortunately, the limitations of hybrid systems in terms of speaker adaptation are well known, since state of the art methods typically used in Gaussian systems (like MLLR) cannot be applied. 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. Then, the resulting speaker adaptation weights are converted into a feature vector form that is used for training speaker models using SVM.

In the present work, we recall the recently proposed Transformation Network features with SVM modelling (TN-SVM) approach and report a set of new experiments on a sub-set of one NIST Speaker Recognition Evaluation 2008 [8] test condition. First, we study the influence of the method used for obtaining the necessary phonetic alignment on the speaker recognition performance. The use of weak automatic transcriptions generated by our automatic speech recognition (ASR) system is evaluated as an alternative to forced alignments of the transcripts provided by NIST. We also explore the feasibility of using a phonetic grammar to obtain the phone sequence needed for network adaptation as an alternative to ASR transcriptions. Second, Nuisance Attribute Projection (NAP) [9] for session varability compensation is applied to the TN-SVM approach.

2. TN features for speaker recognition

2.1. The baseline AUDIMUS hybrid speech recognizer

The AUDIMUS framework developed during the last years of research at INESC ID follows the well known connectionist paradigm [6], allowing the development of several ASR applications, such as the recognition of Broadcast News (BN) for several languages. The 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 combines three MLP outputs trained with Perceptual Linear Prediction features (PLP, 13 static + first derivative), log-RelAtive SpecTrAl features (RASTA, 13 static + first derivative) and Modulation SpectroGram features (MSG, 28 static). When applied to narrowband recordings, the advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives) are also typically used. The number of context input frames is 13 for the PLP, RASTA and ETSI networks and 15 for the MSG network. Although context dependent versions have been developed [10], the baseline system adopted in this work models only monophone units, resulting in MLP networks of 40 softmax outputs for English. The decoder of the recognizer is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition.

Copyright © 2010 ISCA

2.2. MLP/HMM Speaker Adaptation

The Transformation Network [7] (TN) 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. TN adaptation allows keeping unaltered the speaker independent (SI) components of the recognition networks while estimating some sort of speaker dependent (SD) transformation. A block diagram of the TN adaptation approach is depicted in Figure 1.

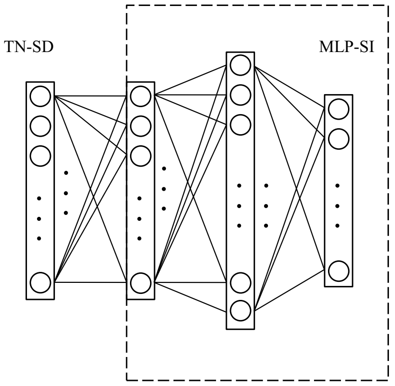

Figure 1: *Block diagram of the Transformation Network (TN) normalization technique.*

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.

2.3. Feature extraction and SVM modelling

Linear speaker dependent mappings for each data segment are independently trained and consequently a speaker adapted transformation matrix is obtained for each segment. A fixed number of training epochs with a relatively small adaptation step is used for estimating the transformation weights.

The transformation mapping is a full square matrix of dimensions $[N_{feat} \cdot N_{context}, N_{feat} \cdot N_{context}]$ where N_{feat} and $N_{context}$ are the size of the feature vector and the number of context input frames of the MLP network respectively. However, it is also possible to estimate tied networks sharing the same weights for all the context frames instead of training a full-matrix, while maintaining a similar speaker adaptation performance. Thus, we reduce the dimensionality of the linear mapping to just $[N_{feat}, N_{feat}]$. In this work, we consider only tied transformation matrices.

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. A linear kernel has been used for training speaker models and min-max normalization in the [0,1] rank has been applied.

2.4. Generation of phone transcriptions

The lack of conversational telephone speech (CTS) orthographically labelled data prevented us from developing an ASR system matched to the characteristics of the NIST Speaker Recognition Evaluation data sets. Consequently, a simple narrow-band speech recognizer with acoustic models trained with down-sampled broadcast news (BN) data was used in this work. The MLP acoustic models were trained on the same 140 hours of manually transcribed HUB-4 speech used for our American English Broadcast News transcription system. We use a 4-gram language model, and a 64k word vocabulary consisting of all the words contained in the HUB-4 training set plus the most frequent words found in the broadcast news texts and Newspapers texts used for LM training. Multiple-pronunciations are allowed and account for a total of 70k entries [11].

This BN ASR system shows relatively good performance in on domain data: in the 1997 evaluation corpus of the NIST HUB-4 American English transcription campaign, the system achieving a 17.6% word error rate (WER). However, informal evaluations allowed us to verify a very weak performance of the narrow-band ASR system trained with downsampled data when it is used for speech recognition of CTS-like data. Word error rates above 70% indicate a strong missmatch of both acoustic, lexical and language models.

Given the need for frame-level phonetic alignments, in our previous work [5], word-level automatic transcriptions provided by NIST were forced aligned using the weak narrow-band system. Then, alignments were used for training the transformation networks.

In the present work, we investigate the impact on the speaker performance of the TN-SVM approach when our own narrow-band speech recognition system is also used for generating the automatic word transcriptions, despite the strong mismatch between the BN training data and the CTS test data.

Alternatively, we also study the possibility of using a simple phone model to produce a phone transcription. In fact, the above mentioned mismatch suggests that using a simple phonetic model could be a feasible alternative for decoding phoneme sequences. Thus, a 3-gram phonetic model was trained using the lexicon of the 64K word vocabulary of the BN system as training data. This phone model was used to obtain frame-level phonetic alignments that are later used for network adaptation.

2.5. NAP for session variability compensation

Session variability compensation for speaker recognition has been recently on the main focus of the research community, partially motivated by the characteristics of the most recent NIST SRE campaigns. As a consequence, some variability compensation methods such as Nuisance Attribute Projection (NAP) [9] have rapidly become part of the current state of the art.

The NAP approach was originally developed for use with SVMs. Thus, it is a compensation method that fits well the characteristics of the proposed TN-SVM technique. NAP aims at removing nuisance attribute-related dimensions in high-dimensional spaces via projections. Basically, these projections expand the orthogonal space to the nuisance directions and can be obtained solving an eigenvalue problem. Applying the trained projections to the high-dimensional vectors allows the reduction of multisession variation for the same speaker and of different channel effects, and increases the “distance” between different speakers.

3. Speaker Recognition Experiments

3.1. Experimental set-up

3.1.1. *Task definition*

Speaker verification is assessed in one sub-set of the *short2-short3* NIST Speaker Recognition Evaluation 2008 test condition [8]. Concretely, we consider the *telephone-telephone* training and test condition.

3.1.2. *Data sets*

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.

Additional training data sets from previous SRE evaluations are used for the development of the speaker recognition system. Concretely, single channel conversation sides of approximately 5 minutes of SRE2004, SRE2005 and SRE2006 evaluations are used for background modelling/training. Multisession 8 conversation sides training data of SRE2004, SRE2005 and SRE2006 are used for training the NAP projection matrix.

3.1.3. *Score calibration*

Every single system is calibrated with the *s-cal* tool available in the Focal toolkit [12]. 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.

3.1.4. *Performance Metrics*

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 DCF point is provided for assessment of the speaker detection systems. Additionally, we also report the Equal Error Rate (EER) and the Detection Error Trade-off (DET) curve for a better evaluation of the speaker recognition systems under study.

3.2. TN-SVM reference system

The TN-SVM reference system is characterized by the use of NIST transcriptions for phonetic alignment generation and no application of any session variability compensation method. The baseline TN-SVM performance compared to two state of the art speaker recognition systems and several other experiments can be found in [5].

3.3. Experiments without NIST transcriptions

The aim of this first set of experiments is to determine the influence of using a weak speech recognition system (*Word model*) or a phonetic decoder (*Phone model*) for generating the phonetic alignment, as an alternative to the use of the transcripts provided by NIST (*Baseline*). DET curves, minDCF and EER scores are reported in Figure 2 and Table 1.

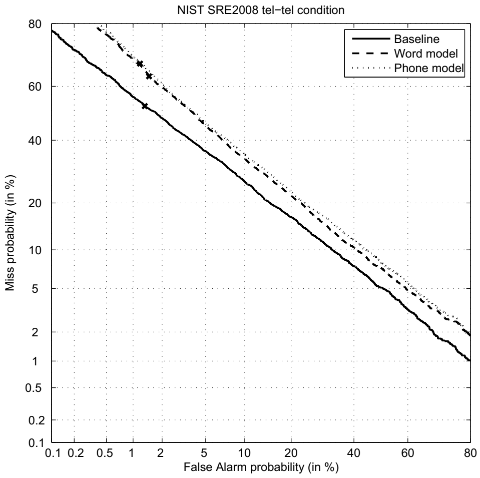

Figure 2: *DET curves of the TN-SVM system evaluating different approaches for generating phonetic alignment: using NIST transcripts (Baseline), using a weak ASR system (Word model) or a phoneme decoder (Phone model).*

System	minDCF (x100)	EER (%)
Phone model	7.97	21.41
Word model	7.82	20.76
Baseline	6.59	17.33

Table 1: *Minimum Detection Cost (x100) and EER.*

A significant performance drop is observed in both cases with respect to the baseline. It is clear that our own ASR produces very low quality transcripts resulting in a relative speaker performance loss of 18.7% and 19.8% in minDCF and EER respectively. The use of a phonetic model obtains slightly worse results than the word model based alternative. This fact confirms the huge mismatch of the ASR system and particularly of the LM to the characteristics of the evaluation data.

On the one hand, high quality word transcripts are of crucial importance for the correct deployment of the TN-SVM approach. On the other hand, it is shown that some speaker dependent information can be obtained even with low quality phonetic alignments. Furthermore, it is important to notice that the weak ASR system is the one that is used in all the three approaches for estimating the speaker transformation matrices. Thus, the availability of a high quality ASR system would have a great benefit not only due to the generation of high quality transcripts, but also to a more accurate estimation of the speaker transforms.

3.4. Experiments with NAP compensation

Figure 3 and Table 2 show the results obtained by applying NAP compensation with different nuisance dimensions to the TN-SVM system. A generalized improved performance of the TN-SVM method with NAP compensation can be observed for all the dimensions of the nuisance space. There are not remarkable performance differences varying the NAP dimensionality, specially at operation points close to the minDCF and the EER. However, it seems that the 16 and 32 dimension systems provide a better performance in the low miss probability region of the DET curve. In fact, it was verified that the decaying slope of the curve formed by the sorted eigenvalues is very slow above the first 50 eigenvalues. This suggests that larger nuisance di-

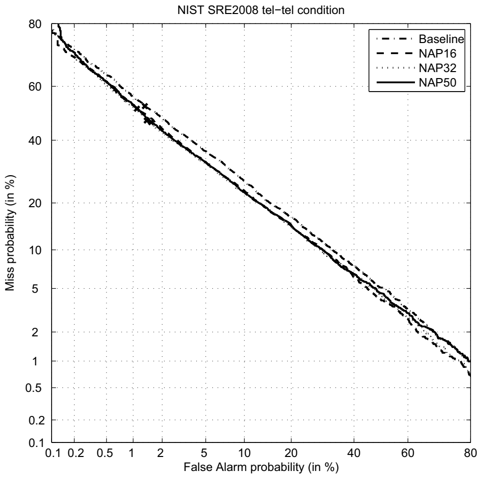

Figure 3: *DET curves of the baseline TN-SVM system compared to the application of NAP variability compensation with different dimensionalities.*

System	minDCF (x100)	EER (%)
NAP 16	6.26	16.26
NAP 32	6.11	16.23
NAP 50	6.20	16.17
Baseline	6.59	17.33

Table 2: *Minimum Detection Cost (x100) and EER .*

mensionalities would not provide additional benefits.

This generalized improvement is not as significant as one could expect. This is very likely due to the fact that cross-channel variability has less influence in this test condition (both training and testing telephone conversations) than in other conditions where room microphones are used together with telephone speech. Thus, it can be said that a moderate degree of channel compensation is achieved (fixed telephone, cellular, etc..), in addition to speaker session variability compensation. In the case of NAP with 32 dimensions, a 7.3% and a 6.4% minDCF and EER relative rate reduction is achieved with respect to the baseline system.

4. Conclusions

Recent advances in speaker recognition tasks comprise the use of features derived from speech recognition adaptation techniques such as MLLR. A novel approach proposed by the authors 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. Previous encouraging results compared to two state of the art baseline detectors motivated the work in this paper.

On the one hand, the importance of automatic speech recognition for phonetic alignment was evaluated and a considerable performance degradation caused by low quality automatic transcriptions was observed. However, it was also shown that a remarkable speaker recognition performance can still be achieved with a poor automatic transcriber or even with a simple phonetic decoder. Our future work plans comprise the use of the phonetic decoder approach to obtain a speaker dependent mapping for any phonetic network independently of the language – both of the speech utterance and of the phone classifier. Thus, similarly to phonotactic approaches applied to the language recognition task, TN features for every single language-dependent phonetic decoder may be obtained resulting in an extended TN vector.

On the other hand, the TN-SVM speaker recognition system benefits from the application of NAP session variability compensation. In the future, the TN features with session variability should be evaluated in a speaker recognition task with stronger cross-channel variability.

5. Acknowledgement

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also by the I-DASH European project (SIP-2007-TP-131703).

6. References

[1] Reynolds, D., Quatieri, T. and Dunn, R., “Speaker verification using adapted Gaussian mixture models”, Digital Signal Processing 10, 19-41, 2000.

[2] Campbell, W. M., Campbell, J. R., Reynolds, D. A., Singer, E. and Torres-Carrasquillo, P. A. “Support vector machines for speaker and language recognition”, Computer Speech and Language, vol. 20, pp. 210-229, 2006.

[3] Ferrer, L., Shriberg, E., Kajarekar, S., Stolcke, A., Sönmez, K., Venkataraman, A. and Bratt, H., “The contribution of cepstral and stylistic features to SRIs 2005 NIST speaker recognition evaluation system”, in Proc. ICASSP 2006, vol. 1, pp. 101-104, Toulouse, 2006.

[4] Stolcke, A., Ferrer, L., Kajarekar, S., Shriberg, E. and Venkataraman, A., “MLLR transforms as features in speaker recognition”, in Proc. Eurospeech 2005, pp. 2425-2428, Lisbon, 2005.

[5] Abad, A. and Luque, J., “Connectionist Transformation Network Features for Speaker Recognition”, in Proc. of Odyssey The Speaker and Language Recognition Workshop 2010, Brno, 2010.

[6] Morgan, N. and Bourlad, H., “An introduction to hybrid HMM/connectionist continuous speech recognition”, IEEE Signal Processing Magazine, vol. 12 (3), pp. 25-42, 1995.

[7] Abrash, V., Franco, H., Sankar, A. and Cohen, M. “Connectionist Speaker Normalization and Adaptation”, in Proc. of Eurospeech 1995, pp. 2183-2186, Madrid, 1995.

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

[9] Solomonoff, A., Campbell, W.M. and Boardman, I., “Advances in channel compensation for SVM speaker recognition”, in Proc. of ICASSP 2005, pp. 629-632, Philadelphia, 2005.

[10] Abad, A. and Neto, J., “Incorporating acoustical modeling of phone transitions in an hybrid ANN/HMM speech recognizer”, in Proc. of Interspeech 2008, pp. 2394-2397, Brisbane, 2008.

[11] Pellegrini, T. and Trancoso, I., “Error detection in automatic transcriptions using Hidden Markov Models”, In Proc. of Language and Technology Conference, 2009.

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