Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Connectionist Transformation Network Features for Speaker Recognition

Alberto Abad[1] *and Jordi Luque*[1,2]

[1] L^2F - Spoken Language Systems Laboratory, INESC-ID Lisboa, Portugal
[2]TALP research center, Universitat Politècnica de Catalunya, Spain

alberto.abad@inesc-id.pt

Abstract

Alternative approaches to conventional short-term cepstral modelling of speaker characteristics have been proposed and successfully incorporated to current state-of-the art systems for speaker recognition. Particularly, the use of adaptation transforms employed in speech recognition systems as features for speaker recognition is one of the most appealing recent proposals. In this paper, we also explore the use of adaptation transform based features for speaker recognition. However, we consider transformation weights derived from adaptation techniques applied to the Multi Layer Perceptrons that form a connectionist speech recognizer, instead of using transforms of Gaussian models. Modelling of the high-dimensionality vectors extracted from the transforms is done with support vector machines (SVM). The proposed method –named Transformation Network features with SVM modelling (TN-SVM)– is assessed and compared to GMM-UBM and Gaussian Super vector systems on a sub-set of NIST SRE 2008. The proposed technique shows promising results and permits further improvements when it is combined with baseline systems.

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, 4]. 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. For instance, cues at the syllable, word or even whole sentence level [4, 5, 6]. An additional limitation of short-term features is that they typically comprise not only the speaker variability information, but they also contain nuisance factors such as channel effects and a strong dependence on the words uttered. These problems enforce the application of feature normalization [7], channel compensation [8] and score normalization methods [9] among others.

In fact, although short-term cepstral based systems are still the core of some of the most successful systems, current state-of-the-art systems typically employ a combination of long-term and short-term features permitting a better characterization of speakers and consequently an improved performance [10, 11, 12].

In [13] 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. Thus, although this approach is also based on cepstral features, unlike standard frame-based cepstral speaker recognition models, it normalizes for the choice of spoken words in text-independent speaker verification. The high-dimensional vectors formed by the transform coefficients are then modelled as speaker features using support vector machines (SVM).

The work that we present in this paper closely resembles the work in [13]. The aim is also to find an alternative approach for speaker recognition consisting on the use of adaptation transforms employed in speech recognition as features for speaker recognition. However, in contrast to [13], 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 [14]. In this case, the use of MLLR or other similar transform methods like constrained MLLR [15] that are employed in Gaussian systems can not be considered any more.

In [16] and [17], several techniques for speaker adaptation of a hybrid ANN/HMM continuous speech recognition system are compared and evaluated. A method referred to as Linear Input Network (LIN) [16] adaptation technique or Transformation Network (TN) [17] employs a trainable linear input network to map the speaker-dependent input vectors to the speaker independent system. The mapping is trained by minimizing the mean squared error of the posterior probabilities at the output of the connectionist system while keeping all the other parameters fixed.

In the present work, TN adaptation weights are used as features for speaker recognition. In this way, this work reports a novel method for incorporating hybrid connectionist speech recognizer based features to speaker recognition tasks. An attractive characteristic of the proposed method –called TN-SVM– is that an independent set of TN features can be obtained for every single feature stream that form the hybrid connectionist speech recognition system. Like in [13], support vector machines are used for speaker modelling of the high dimensionality extracted features. In contrast to the MLLR technique that estimates an affine transformation of the model parameters, the TN adaptation can be seen as a sort of feature pre-processor stage as will be discussed in the next sections.

The paper is organized as follows. In the next section the basic principles of the connectionist speech recognition paradigm are reviewed together with a description of the ASR system for narrow-band data that was built *ad hoc* for this work. Section 3 describes the Transformation Network method for speaker adaptation in connectionist speech recognition systems. Then, the feature extraction processing for speaker recognition and how to build speaker models based on SVM is described.

The experimental assessment of the TN-SVM novel approach and its comparison to a GMM-UBM and a Gaussian Super vector baseline systems on a sub-set of one NIST Speaker Recognition Evaluation 2008 [18] test condition is reported in Section 4. Finally, we present the conclusions in Section 5.

2. Connectionist Speech Recognition

Among the several paradigms proposed during decades of research in automatic speech recognition (ASR), Hidden Markov Models of Gaussian mixtures (HMM/GMM) [19] is doubtless the most widely accepted framework. Alternatively, Artificial Neural Networks (ANN) have also been proposed [20], but despite their high discrimination ability in short-time classification tasks, they have proved inefficient when dealing with long-term speech segments. With the goal of solving the problem of long time modelling of the ANN framework, one of the most successful alternatives to HMM/GMM was later proposed, commonly known as hybrid ANN/HMM or connectionist paradigm [14]. In general, hybrid architectures seek to integrate the ANN ability for estimation of Bayesian posterior probabilities into a classical HMM structure that allows the modelling of long-term speech evolution.

On the one hand, the main advantage of hybrid ANN/HMM are that classification networks are usually considered better pattern classifiers than Gaussian mixtures approaches. Additionally, an appealing characteristic of the hybrid systems is that they are very flexible in terms of merging multiple input streams: the posterior probabilities generated by various networks trained with different streams (usually different feature representations of the same speech data) can be merged, obtaining improved performances. On the other hand, some of the most significant limitations of hybrid systems are related with the lack of flexibility and increased difficulty when context-dependent phone modelling or speaker adaptation is desired, since state of the art methods typically used in Gaussian systems can not be applied.

2.1. Broadcast News Speech Recognizer

The AUDIMUS framework developed during the last years of research at the INESC ID's Spoken Language Systems Laboratory permits the development of several ASR based applications, such as LVCSR of Broadcast News (BN) for several languages [21, 22, 23]. The core speech recognizer of AUDIMUS is a hybrid ANN/HMM system characterized by the use of Multiple Layer Perceptron (MLP) networks that act as phoneme classifiers for estimating the posterior probabilities of a single state Markov chain monophone model. Figure 1 shows a block diagram of the AUDIMUS speech recognizer. The 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). The decoder of the recognizer is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition.

The BN transcription system for American English [23] is based on AUDIMUS architecture, but it incorporates explicit modelling of phone transitions and additional sub-phonetic units in addition to conventional monophone modelling [24]. The MLP acoustic models were trained on 140 hours of manually transcribed HUB-4 speech. The language model is a 4-gram model, with Kneser-Ney modified smoothing, trained on

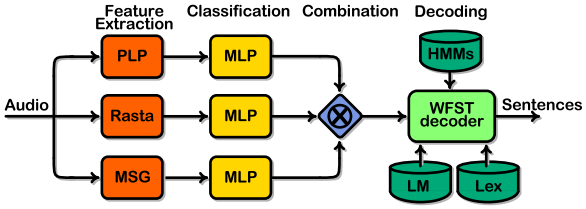

Figure 1: *General block diagram of the AUDIMUS speech recognition architecture.*

150 million words from HUB-4 transcripts, and about 1 billion words of newspaper and news-wire texts. The 64k word vocabulary consists of all the words contained in the HUB- 4 training set plus the most frequent words in the broadcast news texts and Newspapers texts. Multiple-pronunciations are allowed and account for a total of 70k entries. In the 1997 evaluation corpus of the NIST HUB-4 American English transcription campaign the system achieves a 17.6% word error rate (WER).

2.2. Narrow-band Speech Recognizer

The fact that we did not have access to conversational telephone speech (CTS) orthographically labelled data prevented us from having an ASR system in more accordance with the characteristics of the NIST Speaker Recognition Evaluation data sets.

For this work we built a basic *ad hoc* narrow-band speech recognizer with acoustic models trained with down-sampled BN data. That is, we trained phonetic MLP networks for PLP, RASTA, MSG features at 8 KHz sampling rate and we additionally incorporated a new stream with Advanced Font-End from ETSI features (ETSI, 13 static + first and second derivatives) since it was considered adequate for the kind of data. In the narrow-band recognizer only monophone units were modelled. Informal evaluations permitted us to verify a very weak performance of the ASR system for CTS data (word error rates above 70 %). For that reason, it was decided to use our own system for obtaining the phonetic alignment according to the automatic transcriptions provided by NIST, instead of using it also for speech recognition. However, it is worth noting that this weak ASR system is the one that we use for training the speaker adaptation networks.

3. MLP/HMM Speaker Adaptation

In [16] a study of different speaker adaptation techniques applied to hybrid connectionist ANN/HMM systems was presented. The various proposed methods consisted either on the transformation of the parameters (weights) of the ANN component (typically a MLP) and/or augmenting the structure of the speaker independent network. In this work we are interested in an adaptation method able to keep unaltered the speaker independent (SI) components while estimating some sort of speaker dependent (SD) transformation. Thus, an approach known as Transformation Network [17] or Linear Input Network [16] fits in our expectations.

3.1. Transformation Network (TN) normalization

The TN normalization technique employs a trainable linear input network (LIN) to map the SD input vectors to the characteristics of the SI system. It is said to be a normalization method

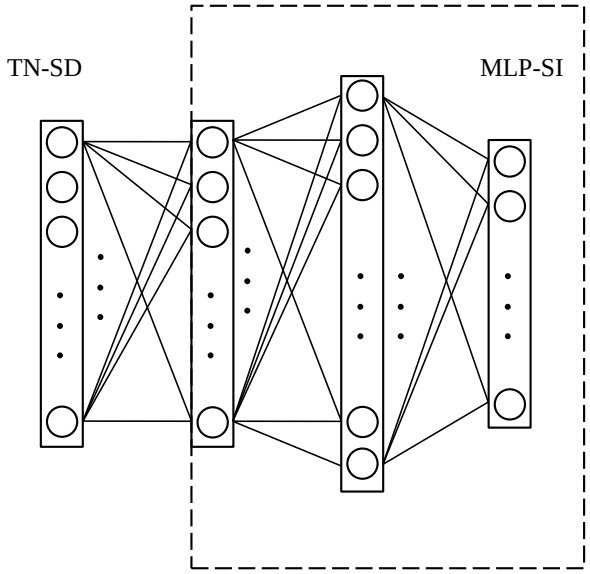

Figure 2: *Block diagram of the Transformation Network (TN) normalization technique. It can be distinguished a MLP speaker independent network (MLP-SI) and a speaker dependent linear input transformation network (TN-SD).*

since its goal is to map the incoming speech to a better representation that enhances the MLP classifier ability to estimate posterior probabilities. A block diagram of the TN adaptation approach is depicted in Figure 2

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. That is, the input mapping is trained with the enrolment data available for a given speaker by minimizing the mean square error of the probabilities at the output of the connectionist system while keeping all the other parameters fixed. 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, according to [16] the TN approach is architecture dependent, hence it can not be considered a generalized spectral normalization technique.

3.2. Feature extraction

Phonetic alignments of every speech segment with the automatic transcription provided by NIST are generated with the narrow-band speech recognition system described in section 2.2.

Then, the frame-level forced alignments are used to train the linear speaker dependent mapping for each data segment independently. 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.

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. For instance, in the case of the PLP network the number of transformation weights that form the adaptation matrix is 114244 ($[26 \cdot 13, 26 \cdot 13]$). In some preliminary experiments, we confirmed that it is 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 can reduce the dimensionality of the linear mapping to just $[N_{feat}, N_{feat}]$. In this work, we consider only tied transformation matrices.

One characteristic of the hybrid systems is the normalization that is applied to the MLP inputs. Each input feature is normalized to have zero mean and unit variance according to estimates obtained during training. It is known [16] that the speaker adaptation performance is enhanced by estimating new feature normalization from the adaptation data. Hence, in this work we also estimate mean and variance normalization feature statistics in a per segment basis. In fact, it is expected that the statistics are related to the adaptation weights and that provide additional speaker dependent information.

The coefficients from the linear mapping obtained for each speaker are concatenated in a vector. An independent TN feature vector can be obtained for each network stream: PLP, RASTA, MSG and ETSI. Except as otherwise noted, mean and variance of the features extracted from the speaker segment are also incorporated to the feature vector.

3.3. SVM modelling

Connectionist transformation network feature vectors are used to train SVM target speaker models: the target speaker feature vector is used as positive example, while the feature vectors extracted from a background data set (in the same fashion described above) are used as negative examples.

The dense implementation of the libSVM toolkit [25] is used for training. In this work, linear kernel has been used for training speaker models. Like in [13], the dynamic ranges of the feature vector components are normalized in order to reduce sensitive of the SVM kernel function to the magnitude of the feature values. In this case, we apply a min-max normalization in the [0,1] rank.

4. Speaker Recognition Experiments

4.1. Experimental set-up

4.1.1. Task definition

Speaker verification consists of determining whether a specified speaker is speaking during a given segment of speech. In this work, it is assessed in one sub-set of the *short2-short3* NIST Speaker Recognition Evaluation 2008 test condition [18]. Concretely, we consider the *telephone-telephone* training and test condition.

4.1.2. Data sets

Training and testing data sets of the *telephone-telephone* condition consists of one 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 development of the speaker recognition systems. Concretely, single channel conversation sides of approximately 5 minutes of SRE2004, SRE2005 and SRE2006 evaluations are used for background modelling/training and score normalization.

Automatically produced English language word transcripts are available for all speech segments with word error rates in the range of 15-30%. These automatic transcripts are used for phonetic alignment needed for Transformation Network feature extraction.

4.1.3. Baseline systems

Two well-known state of the art methods – without session variability compensation techniques applied – have been developed for comparison.

- ▷ GMM-UBM

  Gaussian Mixture Models (GMM) for each target speaker are obtained with MAP adaptation of the Gaussian means of a Universal Background Model (UBM) [1]. The GMM-UBM feature front-end extracts 19 PLP static features with log-RASTA processing and the frame energy from a sliding window of 20 milliseconds with a step size of 10 milliseconds. First and second derivatives are concatenated to form 60 element feature vectors. A well-trained MLP speech-non-speech detector is combined with the alignment generated by a simple bi-Gaussian model of the log energy distribution computed for each speech segment to remove low-energy and high likely non-speech frames. Finally, mean and variance feature normalization is performed in a segment-based fashion. Gender dependent UBMs of 1024 components are trained with 742 female and 520 male speaker segments obtained from SRE2004 and SRE2005 training data sets. UBM means are adapted with 5 MAP iterations with a relevance factor of 16 to obtain the speaker models. Raw log-likelihood scores are T-normalized using 200 speech segments (100 for female and 100 for male) randomly selected from the background SRE2004 and SRE2005 data set.

- ▷ GSV-SVM

  The Gaussian Super Vector (GSV) system concatenates the mixture means of the MAP adapted Gaussian speaker models trained in the GMM-UBM system to obtain super vectors of every speech segment. The linear SVM kernel of [3] is used for training the speaker models with the libSVM tool. Min-max rank normalization is also applied to the super vectors. For training, the target speaker super vector is used as the positive example, while a set of background super vectors are used as negative samples. The background set used as negative examples for SVM training is formed by 493 female super vectors (293 SRE2004 and SRE2005 background segments + 200 SRE2006 additional segments) and 453 male speech segments (254 SRE2004 and SRE2005 background segments + 199 SRE2006 additional segments). It was verified that score normalization strategies do not provide a remarkable improvement in discriminative scoring systems like the GSV-SVM, hence we decided not to apply score normalization.

4.1.4. Score calibration and fusion

Every single system is calibrated with the *s-cal* tool available in the Focal toolkit [26]. It permits 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. In the following experiments, whenever more than one system is combined, the corresponding scores are fused together at this stage. Thus, the linear logistic regression permits simultaneous calibration and fusion of multiple systems scores. All calibration and fusion parameters are gender-dependent. A five-fold cross-validation strategy on the test set is applied to simultaneously estimate the calibration and fusion parameters and to evaluate speaker detection systems.

4.1.5. Performance Metrics

The detection cost function (DCF) is the metric used in NIST evaluations and it is defined as a weighted sum of miss and false alarm error probabilities: $C_{det} = C_{Miss} \cdot P_{Miss|Target} \cdot P_{Target} + C_{FalseAlarm} \cdot P_{FalseAlarm|NonTarget} \cdot (1 - P_{Target})$. The parameters of this cost function are the relative costs of detection errors, C_{miss} and $C_{FalseAlarm}$, and the a priori probability of the specified target speaker, P_{target}. The parameter values are the ones used in NIST 2008, that is, P_{target}=0.01, C_{miss}=10, $C_{FalseAlarm}$=1. In this work, we provide the minimum DCF point 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.

4.2. Experimental results

4.2.1. Network selection experiments

The aim of this first set of experiments is to validate the Transformation Network features trained with Support Vector Machines (TN-SVM) approach and to verify that TN features effectively contain speaker information that can be used for speaker recognition applications. It is worth noting that in preliminary experiments we observed insignificant improvements in TN-SVM systems with score normalization strategies, hence it has not been applied in any of the following TN-SVM detection systems.

As commented previously, one of the attractive properties of connectionist ASR systems is the possibility of having multiple streams and the flexibility for merging them. The fact that our narrow-band speech recognizer uses 4 different networks permits us extracting 4 independent TN feature vectors. In order to validate our proposal, we have first trained single TN-SVM detectors with the features extracted from the four different stream MLPs. DET curves, minDCF and EER scores are reported in Figure 3 and Table 1. The first remarkable observation is that that every single system has the ability for speaker discrimination. The best single TN-SVM detector is the one based on the ETSI network. One possible reason may be the higher dimensionality of the TN feature vector. However, the

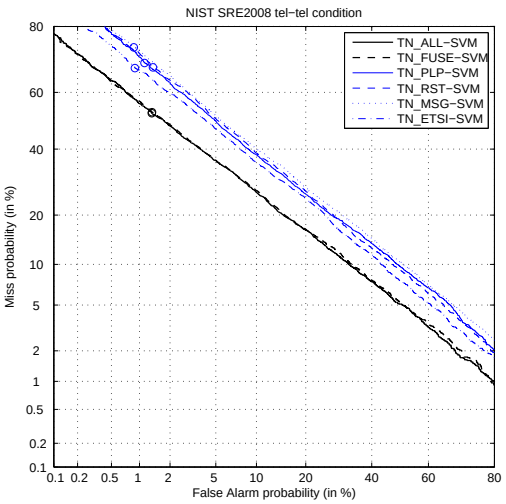

Figure 3: *DET curves of each individual stream based detector (TN_PLP-SVM, TN_RST-SVM, TN_MSG-SVM and TN_ETSI-SVM), of the fusion of the four systems (TN_FUSE-SVM) and of the four concatenated feature vectors system (TN_ALL-SVM).*

System	minDCF (x100)	EER (%)	#feats
TN_PLP-SVM	8.08	22.69	728
TN_RST-SVM	8.19	23.07	728
TN_MSG-SVM	8.30	23.65	840
TN_ETSI-SVM	7.70	21.96	1599
TN_FUSE-SVM	6.63	17.55	—
TN_ALL-SVM	6.59	17.33	3895

Table 1: *Minimum Detection Cost (x100), EER and number of features of each individual stream based detector (TN_PLP-SVM, TN_RST-SVM, TN_MSG-SVM and TN_ETSI-SVM), of the fusion of the four systems (TN_FUSE-SVM) and of the four concatenated feature vectors system (TN_ALL-SVM).*

MSG based system shows the weaker detection results and it is of slightly higher dimensionality than the PLP and RASTA based TN features. In fact, it seems that ETSI based features perform better not only because of the TN vector dimension, but also because the ETSI features are better suited to the characteristics of the speech data involved in this evaluation.

The availability of multiple TN-SVM systems opens the possibility for merging or fusion. The first approach attempted was the fusion of the four detectors at the score level (TN_FUSE-SVM). Another possibility is to form a high dimensionality vector concatenating the TN features extracted from the four individual MLP networks (TN_ALL-SVM or simply TN-SVM). In both cases, a considerable improvement with respect to the best individual system is achieved. It seems that using a single feature vector of high dimensionality achieves slightly better results than the fusion at the score level in both minDCF and EER score, although the DET curve shows very similar results at the different possible operation points. In the case of the TN_ALL-SVM detector, a 14.4% and a 21.1% minDCF and EER relative rate reduction is achieved with respect to the best individual system (TN_ETSI-SVM).

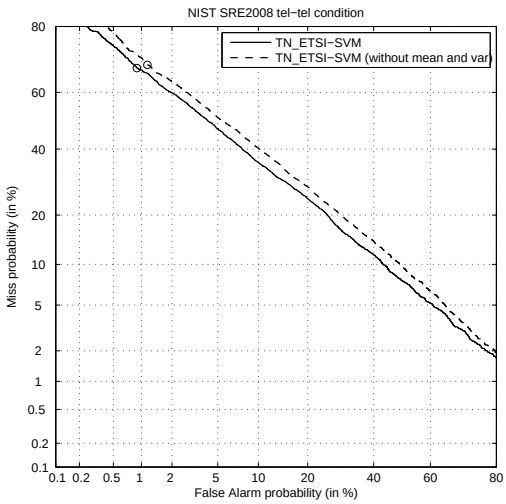

Figure 4: *DET curves of the single stream TN_ETSI-SVM detector with and without feature mean and variance statistic features.*

System	minDCF (x100)	EER (%)	#feats
TN_ETSI *wmv*	7.70	21.96	1599
TN_ETSI *womv*	8.05	23.66	1521

Table 2: *Minimum Detection Cost (x100), EER and number of features of the single stream TN_ETSI-SVM detector with and without feature mean and variance statistic features.*

4.2.2. Mean and variance feature experiments

The use of mean and variance statistics computed from the speaker specific data used for feature normalization provides improvements of the speaker adaptation techniques used in connectionists systems for ASR. According to this, it was decided to concatenate mean and variance feature statistics of the speaker data to the transformation matrix weights to form the TN vector. In Figure 4 and Table 2 are compared the best single stream performing system (TN_ETSI-SVM) when the feature vector incorporates mean and variance statistics (with mean and variance $\equiv$ *wmv*) and when it does not (without mean and variance $\equiv$ *womv*). Regarding these results, it can be confirmed the importance of the features mean and variance also for speaker recognition with TN features.

4.2.3. Baseline systems comparison

The GMM-UBM and the GSV-SVM baseline systems are compared to the proposed TN-SVM technique in Figure 5 and Table 3. The proposed TN-SVM detector clearly outperforms the state-of-the-art GMM-UBM detector with t-norm score normalization in terms of minimum detection cost. However, it is worth noting that the differences between these two detectors strengthens as long as we approximate to the EER operation point and are almost equivalent in other points of the DET curve. It is not clear the reason for these differences at the different detection thresholds. With respect to the GSV-SVM baseline technique, slightly worse results in terms of minimum

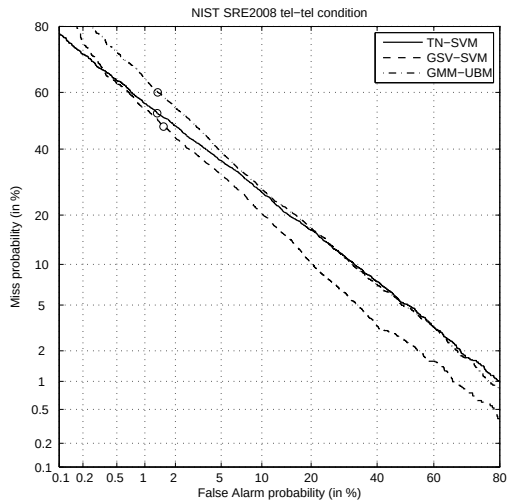

Figure 5: *DET curves of the TN-SVM proposed system and of the GMM-UBM (with t-norm) and GSV-SVM baseline systems.*

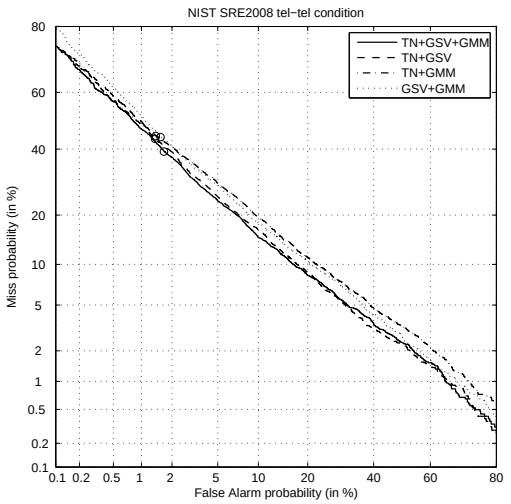

Figure 6: *DET curves of the TN-SVM proposed system when it is fused with the GMM-UBM system (TN+GMM), with the GSV-SVM system (TN+GSV) and with both (TN+GSV+GMM).*

System	minDCF (x100)	EER (%)
GMM-UBM	7.34	17.88
GSV-SVM	6.32	14.65
TN-SVM	6.59	17.33

Table 3: *Minimum Detection Cost (x100) and EER of the TN-SVM proposed system and of the GMM-UBM (with t-norm) and GSV-SVM baseline systems.*

System	minDCF (x100)	EER (%)
GMM+GSV	5.96	13.87
TN+GMM	5.84	14.45
TN+GSV	5.72	12.95
TN+GSV+GMM	5.58	12.40

Table 4: *Minimum Detection Cost (x100) and EER of the fusion of the two baseline systems (GMM+GSV) and of the TN-SVM proposed system when it is fused with the GMM-UBM system (TN+GMM), with the GSV-SVM system (TN+GSV) and with both (TN+GSV+GMM).*

detection cost are achieved. Once again, the TN-SVM detector seems to perform worse with respect to the GSV-SVM system as long as we move away from the threshold of minimum detection cost. Anyway, it can be stated that the TN-SVM is able to provide comparable speaker detection capabilities to two state of the art speaker recognition systems.

4.2.4. Systems combination

Finally, the last set of experiments is focused on the study of the ability of the new speaker detector to fuse with other baseline systems. Results shown in Figure 6 and Table 4 demonstrate that the TN-SVM detector permits a considerable improvement when it is fused. For instance, when it is fused with the GSV-SVM system a 9.5 % and a 11.6 % minDCF and EER relative reduction is achieved with respect to the GSV-SVM (that was the best baseline detector). A considerable improvement can also be observed with respect to the fusion of the two baseline systems (GMM+GSV) when the three detectors are fused (TN+GSV+GMM). Regarding these results, we can conclude that the TN approach provides complementary cues for speaker recognition.

4.3. Discussion

In this paper we have shown how features derived from ASR adaptation techniques used in connectionist ANN/HMM speech recognition systems can be used for speaker recognition task. The Transformation Network features technique for speaker recognition proposed in this work has shown relative good performance when compared to other state of the art baseline detectors. Although the method is based on short-time cepstral feature extraction, the transformation features are estimated for a whole speech segment. Consequently, the extracted features are normalized independently on the words uttered. Moreover, the proposed approach seems particularly convenient for fusion with more conventional acoustic model-based approaches, providing complementary discrimination abilities.

Some questions may arise related to the proposed technique and the influence of the ASR system employed. First, a very weak speech recognition system has been used for both phonetic alignment and network adaptation. In fact, we suspect that using a more robust system would permit an improved estimation of the transformation network, and consequently, the extraction of more informative speaker recognition features that would allow an improved detection performance. Second, the use of automatic transcriptions generated by our own speech recognizer has not been experienced and only transcripts provided by NIST have been used. The use of weak transcriptions as the ones provided by our system should be evaluated in the future, testing its impact on the speaker recognition performance. This fact arises an interesting related problem: whether it is feasible to apply the TN-SVM method in a way that it is

not necessary an automatic transcription, for instance using an open loop grammar of phonetic units or similarly to [15]. This is a very interesting question that remains open for future investigation.

Another possibility for future research is related to the adaptation method considered and the way it is applied. The use of the Transformation Network approach has shown the ability for capturing speaker cues, however it is likely that other methods applied in connectionist systems can be also useful. Additionally, some adaptation parameters such as the number of adaptation epochs and the size of the adaptation step have been determined heuristically based on few samples. It is likely that we can obtain better results using other adaptation strategies, for instance using cross-validation data for determining an optimum number of adaptation epochs.

Some implementation aspects may have had a considerable influence on the performance achieved by the proposed method and the other baseline systems. Particularly, a relatively small background set has been used as negative examples for SVM model training or for background modelling. Thus, an improved detection performance of the TN-SVM method can be expected with a larger background data set.

Finally, it is worth noticing that session variability compensation techniques have not been applied neither to the baseline techniques nor to the proposed approach. The application of such techniques and the benefit that might provide is a subject of future research.

5. Conclusions

Recent advances in speaker recognition tasks comprise the use of features derived from speech recognition adaptation techniques such as MLLR. In this paper, we have shown that is is also possible to extract meaningful features for speaker recognition derived from adaptation techniques used in connectionist ANN/HMM speech recognition systems. Concretely, we have adapted speaker independent MLP networks with a method known as Transformation Network in order to obtain a set of speaker dependent linear mappings. The weights of these mappings are concatenated to form a high-dimensionality feature vector that is used for training speaker SVM models. This novel approach is named TN-SVM. The proposed TN-SVM speaker recognition system has been evaluated in a sub-set of the NIST SRE 2008 core condition showing detection performance comparable to two state of the art baseline detectors. Additionally, the novel method has probed to be adequate for combination with other conventional baseline systems due to the complementary speaker cues that provides.

6. Acknowledgement

This work was partially funded by the I-DASH European project (SIP-2007-TP-131703). The authors would like to thank to Prof. Isabel Trancoso for her support.

7. 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] Campbell, W. M., Sturim, D. E. and Reynolds, D. A., "Support vector machines using GMM supervectors for speaker verification" IEEE Signal Processing Letters, vol. 13(5), pp. 308-311, 2006.

[4] 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 Proceedings ICASSP 2006, vol. 1, pp. 101-104, Toulouse, 2006.

[5] Ferrer, L., Bratt, H., Gadde, V.R.R., Kajarekar, S., Shriberg, E., Snmez, K., Stolcke, A. and Venkataraman, A., "Modeling duration patterns for speaker recognition", in Proceedings of the Eurospeech 2003, pp. 2017-2020, Geneva, 2003.

[6] Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A. and Stolcke, A. "Modeling Prosodic Feature Sequences for Speaker Recognition", Speech Communication, Vol. 46(3-4), pp. 455-472, 2005.

[7] Pelecanos, J. and Sridharan, S., "Feature warping for robust speaker verication", in Proceeding of Odyssey Speaker and Language Recognition Workshop 2001, Crete, 2001.

[8] Kenny, P., Boulianne, G. Ouellet, P. and Dumouchel, P., "Joint factor analysis versus eigenchannels in speaker recognition", IEEE Trans. Audio, Speech, Lang. Process., vol. 15(4), 2007.

[9] Auckenthaler, R., Carey, M. and Lloyd-Thomas, H., "Score Normalization for Text-Independent Speaker Verification Systems", Digital Signal Processing, vol. 10, pp. 42-54, 2000.

[10] Sturim, D. E., Campbell, W. M., Karam, Z. N., Reynolds, D.A. and Richardson, F. S., "The MIT Lincoln Laboratory 2008 Speaker Recognition System", in Proceedings Interspeech 2009, pp. 2359-2362, Brighton, 2009.

[11] Kajarekar, S., Scheffer, N., Graciarena, M., Shriberg, E., Stolcke, A., Ferrer, L. and Bocklet, T., "The SRI NIST 2008 Speaker Recognition Evaluation System", in Proceedings of ICASSP 2009, Taipei, Taiwan, 2009.

[12] Burget, L., Fapšo, M., Hubeika, V., Glembek, O., Karaát, M., Kockmann, M., Matějka, P., Schwarz, P. and Ćernocký, J., "BUT system for NIST 2008 speaker recognition evaluation", in Proceedings of Interspeech 2009, pp. 2335-2338, Brighton, 2009.

[13] 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.

[14] 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.

[15] Ferràs, M., Leung, C. C., Barras, C. and Gauvain, J-L, "MLLR Techniques for Speaker Recognition", in Proceedings of Odyssey Speaker and Language Recognition Workshop 2008, Stellenbosch, 2008.

[16] Neto, J., Almeida, L., Hochberg, M., Martins, C., Nunes, L., Renals, S., Robinson,T., "Speaker adaptation for hybrid HMM-ANN continuous speech recognition system", in Proceedings of Eurospeech 1995, pp. 2171-2174, Madrid, 1995.

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

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

[19] Rabiner, L.R., “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, vol. 77(2), pp. 257-286, 1989.

[20] Lippmann, R.P., “Review of neural networks for speech recognition”, Neural Computation, vol. 1(1), pp. 1-38, 1990.

[21] Meinedo, H., Caseiro, D., Neto, J. and Trancoso, I., “AUDIMUS.media: A Broadcast News speech recognition system for the European Portuguese language”, in Proc. Int. Conf. of Computational Processing of Portuguese Language (PROPOR), 2003.

[22] Martinez, R., Neto, J. and Caseiro, D., “Statistical machine translation of broadcast news from Spanish to Portuguese”, in Proc. Int. Conf. of Computational Processing of Portuguese Language (PROPOR), 2008.

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

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

[25] Chang, C.-C. and Lin, C-J, “LIBSVM - A Library for Support Vector Machines”, URL: http://www.csie.ntu.edu.tw/ cjlin/libsvm/index.html.

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