ELSEVIER

Speech Communication 31 (2000) 181–192

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Robustness to telephone handset distortion in speaker recognition by discriminative feature design

Larry P. Heck [a,*], Yochai Konig [b], M. Kemal Sönmez [c], Mitch Weintraub [a]

[a] *Nuance Communications, 1380 Willow Road, Menlo Park, CA 94025, USA*
[b] *Utopy Incorporated, 330 Fell Street, San Fransisco, CA 94102, USA*
[c] *SRI International, Menlo Park, CA 94025, USA*

Received 30 October 1998; received in revised form 30 September 1999

Abstract

A method is described for designing speaker recognition features that are robust to telephone handset distortion. The approach transforms features such as mel-cepstral features, log spectrum, and prosody-based features with a non-linear artificial neural network. The neural network is discriminatively trained to maximize speaker recognition performance specifically in the setting of telephone handset mismatch between training and testing. The algorithm requires neither stereo recordings of speech during training nor manual labeling of handset types either in training or testing. Results on the 1998 National Institute of Standards and Technology (NIST) Speaker Recognition Evaluation corpus show relative improvements as high as 28% for the new multilayered perceptron (MLP)-based features as compared to a standard mel-cepstral feature set with cepstral mean subtraction (CMS) and handset-dependent normalizing impostor models.

Zusammenfassung

Der Artikel beschreibt eine Methode zur Bestimmung von Merkmalen zur Sprechererkennung, die robust gegen Verzerrung durch den Telephonhörer sind. In unserem Verfahren werden die Merkmale, wie z.B. Mel-Cepstrum, logarithmisches Spektrum, oder prosodische Merkmale, durch ein nicht-lineares künstliches neuronales Netz transformiert. Das neuronale Netz wird diskriminativ darauf trainiert, die Sprechererkennungsrate bei unterschiedlichen Telephonhörern im Training und Test zu maximieren. Der Algorithmus braucht weder Stereo-Sprachaufzeichnungen im Training, noch bedarf er manueller Feststellung des Hörertyps im Training oder Test. Die Ergebnisse auf dem 1998 NIST Sprechererkennungskorpus zeigen eine relative Verbesserung von bis zu 28% durch die neuen neuronalen-Netz-Merkale, verglichen mit gewöhnlichen Mel-Cepstrum-Merkmalen, Subtraktion der Cepstrum-Mittelwerte und hörerspezifischen normalisierenden Impostor-Modellen.

Résumé

Une méthode est décrite pour l'extraction de vecteurs de caractéristiques robustes aux distortions provenant du type de téléphone utilisé dans des applications de reconnaissance du locuteur. La technique transforme les vecteurs de caractéristiques tels que le Mel-cepstre, le log-spectre et les caractéristiques basées sur la prosodie, à l'aide de réseau de neurones non-linéaire. Le réseau de neurones est entraîné de manière discriminante pour maximiser la performance du

* Corresponding author. Tel.: +1-650-874-7746; fax: +1-650-847-7878.
E-mail address: heck@nuance.com (L.P. Heck).

0167-6393/00/$ - see front matter
PII: S0167-6393(99)00077-1

système de reconnaissance du locuteur, spécifiquement dans des conditions où des types de téléphone différents sont utilisés lors de l'entraînement et de la vérification. L'algorithme ne requiert, ni enregistrement stéréo de la session d'entraînement, ni étiquettage manuel des types de téléphone utilisés à l'entraînement et à la vérification. Les résultats sur le corpus *1998 NIST Speaker Recognition Evaluation* montrent une amélioration relative atteignant 28% avec les nouvelles caractéristiques basées sur le réseau de neurones. Le système de référence utilise des vecteurs de caractéristiques basés le MEL-cepstre avec soustraction du cepstre moyen ainsi que des modèles d'imposteurs dépendant du type de téléphone.

Keywords: Speaker recognition; Speaker verification; Speaker identification; Channel compensation; Channel robustness; Telephone handset distortion; Feature extraction; Neural network; Discriminative design

1. Introduction

A dominant source of errors in telephone-based speaker recognition systems is the distortion of the speech signal caused by the microphone in the telephone handset (e.g., electret, carbon-button). The distortion can cause an order-of-magnitude increase in speaker recognition error rates when verification tests are completed on a handset type that does not match the enrollment handset type, even after standard channel compensation techniques are applied (Reynolds, 1995; Heck and Weintraub, 1997). Given that verification tests with mismatched telephone handsets occur frequently in practice, handset distortion poses a significant barrier to successful deployment of the technology.

Previous handset and channel compensation approaches can be grouped into three broad classes: (1) model-based, (2) score-based, and (3) feature-based. Model-based compensation methods for speaker recognition include an approach (Murthy et al., 1999) that transforms speaker model variances based on stereo recordings across multiple handsets. A single transform is estimated with a development set of speakers, and is applied during enrollment of all new speakers. The transform is built to be independent of the telephone handset used during enrollment. In contrast, another model compensation method (Heck and Weintraub, 1997) explicitly utilized a handset classifier to select handset-specific, speaker-independent normalizing models for each speaker. The handset-type of the normalizing model was the same handset-type used during the enrollment of the speaker's model. The use of handset-dependent normalizing models significantly reduced error rates over a state-of-the-art robust speaker verification system, which used both cepstral mean subtraction (CMS) and a normalizing world cohort model.

A score-based handset and channel compensation method for speaker recognition systems called HNORM was presented in (Reynolds, 1997b). As with the model-based compensation algorithm described above, the method utilized an automatic handset detector during the training of the speaker model. However, this approach also used the handset detector to classify the test utterance and utilized a database of speech utterances from a representative set of impostor speakers that were labeled according to the type of handset used during the recording. Speech utterances from each type of handset were scored against the speaker model, and the score distribution was modeled with a single Gaussian. The compensation consists of normalizing the test utterance score by removing the handset-dependent bias and scaling (mean and standard deviation) of the impostor score distribution.

While model- and score-based methods have received recent attention in the literature, most of the past work on handset and channel compensation methods for speaker recognition have focused on feature-based methods. Feature extraction plays an important role in speaker recognition where the objective is to extract and select features that provide speaker discrimination while being invariant to non-speaker-related conditions such as handset type, sentence content, and channel effects. Although cepstral-based features are widely used in the field, their design criterion is not

consistent with the objective of maximizing speaker recognition rates. As a result, significant research has been devoted to designing new feature types for robust speaker and speech recognition.

CMS (Furui, 1981) and RASTA-PLP (Hermansk, 1991) are two of the more standard feature-based compensation techniques used to provide robustness to channel effects. However, it is well known that handset and channel mismatches can still be a significant source of errors after CMS or RASTA-PLP (NIST, 1996, 1997, 1998). For this reason, more sophisticated cepstrum transformation methods have been proposed in the literature. In (Stern et al., 1994; Liu et al., 1994; Neumeyer and Weintraub, 1994), cepstral compensation vectors were derived from a stereo database and applied to the training data to adjust for environmental changes. The compensation vectors depend either on the signal-to-noise ratio (SNR) or on the phonetic identity of the frames. In (Mammone et al., 1996), an affine transformation of the cepstral vectors was estimated from a stereo portion of the database under study, and then applied to the training data. In (Murthy et al., 1999), we introduced a new filterbank design and spectral slope-based features to minimize the effects of telephone handset and channel distortions on speaker identification performance.

In recent work by Quatieri et al. (1998), a feature-based compensation method was developed to specifically treat the land-line telephone handset mismatch problem between electret and carbon-button. A one-way nonlinear mapper was designed by matching the spectral magnitude of the distorted signal (carbon-button handset) to the output of a nonlinear channel model driven by an undistorted reference (electret handset). The mapper was trained with stereo recordings of utterances over a small number of handsets in HTIMIT (Reynolds, 1997a). The mapper consisted of a polynomial nonlinearity combined with a linear pre- and post-filter trained to minimize the mean-squared spectral magnitude error using a gradient descent technique.

Discriminative feature design approaches have been developed that use an objective function directly related to classification performance (rather than representational performance). These discriminative feature design techniques have been studied mainly for the speech recognition task (Bengio et al., 1992; Chengalvarayan and Deng, 1997; Euler, 1995; Paliwal et al., 1995). Bengio and his colleagues suggested a global optimization of a combined multilayered perceptron (MLP)-hidden Markov model (HMM) speech recognition system with the maximum mutual information (MMI) criterion, where the outputs of the neural network constituted the observation sequence for the HMM (Bengio et al., 1992). Euler (1995) reported improved HMM speech recognition performance on spelled names when employing a discriminative training approach for designing a feature-based transformation matrix. A recent extension of this work focused on the use of a parallel network of nonlinear and linear feature mappings (Rahim et al., 1997). The linear mapping was initialized to produce standard cepstral-based features. Output feature vectors from the nonlinear neural network were added to the linear cepstral feature vectors, resulting in a single modified feature vector that was fed into the HMM classifier. This approach in effect used the nonlinear neural network-based feature extractor to "correct" the standard cepstral-based features so that the resulting feature set was more robust to channel distortions.

In this paper, we develop a discriminative feature design approach for speaker recognition. Our approach specifically focuses on the problem of telephone handset mismatch between training and testing. As compared to previous speaker recognition feature design efforts, our training procedure directly maximizes speaker recognition performance, does not require stereo recordings of speech across multiple handset types, and does not require manual labeling of the handset types in either training or testing. The new features have been used successfully for speaker verification, and have shown significant improvements in performance over all handset training–testing combinations in the 1998 Speaker Recognition Evaluation coordinated by the National Institute of Standards and Technology (NIST, 1996, 1997, 1998).

We begin this paper by defining the system architecture in Section 2 followed by a description of the proposed feature-based handset compensation

method in Section 3. The development database and experimental results on the 1998 NIST evaluation set are described in Section 4. Finally, we conclude and describe directions for future work in Section 5.

2. System architecture

A general block diagram of the proposed system for discriminative feature design is shown in Fig. 1. The speech signal contains information about the speaker's identity and the content of the spoken sentence. For speech recorded on the telephone, the signal will also be contaminated by noise, be bandlimited, and be distorted by the transducer in the telephone handset.

The feature extraction is composed of two parts: an initial feature analysis and a nonlinear feature transformation. The feature analysis is used to convert the speech signal into a collection of feature vectors such as log spectrum or cepstrum. These features are then processed by the nonlinear feature transformation before being passed on to the speaker recognition classifier. The feature transformation is implemented as an MLP-based artificial neural network.

During the feature design phase, the speaker recognition classifier is also implemented as a MLP-based neural network. Like the feature transformation component, the classifier is trained to reduce the effects of nonlinear handset distortions on speaker discrimination. However, after the feature design phase, other classifier types can be used to complete the speaker recognition task. For the experiments we describe in this paper, we used a state-of-the-art text-independent speaker recognition classifier based on a Bayesian-adapted Gaussian mixture model (GMM) (Reynolds, 1997b). The framework for the discriminative feature design phase is described next.

3. Feature-based handset compensation method

3.1. Cross entropy cost function

Let $X = \{X_1, X_2, \ldots, X_T\}$ be a sequence of feature vectors belonging to the speaker class C_i. We seek to maximize the speaker recognition performance and robustness by minimizing the cross entropy cost function (Baum and Wilczek, 1988)

$$J = -E\left\{ \sum_{i=1}^{M} [d_i \log Y_i(\mathscr{F}(X, \Psi); \Lambda) + (1 - d_i) \log(1 - Y_i(\mathscr{F}(X, \Psi); \Lambda))] \right\}, \tag{1}$$

where $E\{\cdot\}$ is the expectation over the dataset, d_i the desired speaker decision, $\mathscr{F}(X, \Psi)$ the mapping of the input features X with the corresponding set of parameters Ψ, Λ are the parameters of the classifier, and $Y_i(\mathscr{F}(X, \Psi); \Lambda)$ is the ith output of the speaker recognition system. Minimization of

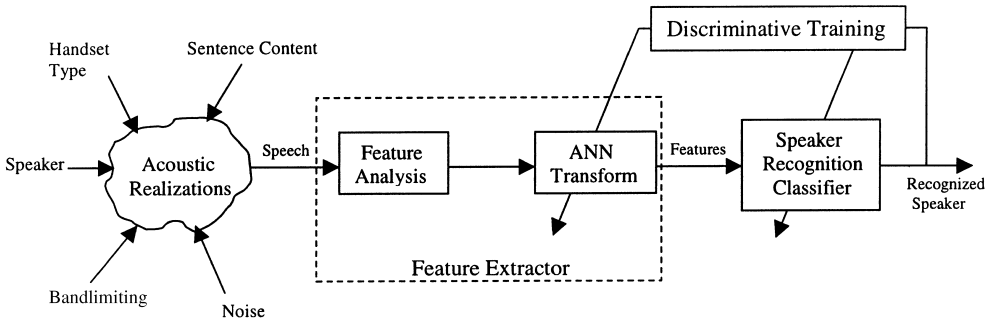

Fig. 1. Block diagram of discriminative feature and classifier design approach. The speech signal is corrupted by a number of environmental factors, which the approach attempts to compensate for by adapting the artificial neural network (ANN) feature transform and speaker recognition classifier based on an estimate of speaker recognition performance.

the cross entropy cost function in this work is achieved by jointly optimizing the parameters of the feature extractor (Ψ) and classifier (Λ).

The cross entropy function has many properties that make it an attractive cost function to use in the design of the feature mapping. First, when the system parameters are chosen to minimize Eq. (1), the outputs estimate Bayesian a posteriori probabilities (Richard and Lippmann, 1991). This property gives an intuitive interpretation of the outputs, and facilitates the straightforward combination of multiple systems of this type for higher-level decision making. Second, maximizing the a posteriori probabilities of the speakers leads to maximization of the speaker classification performance.

To minimize the cross entropy cost function in Eq. (1), we use the standard back-propagation algorithm (Rumelhart et al., 1986). Minimizing the cross entropy cost function can be interpreted as minimizing the Kullback–Liebler probability distance measure or maximizing mutual information (Baum and Wilczek, 1988).

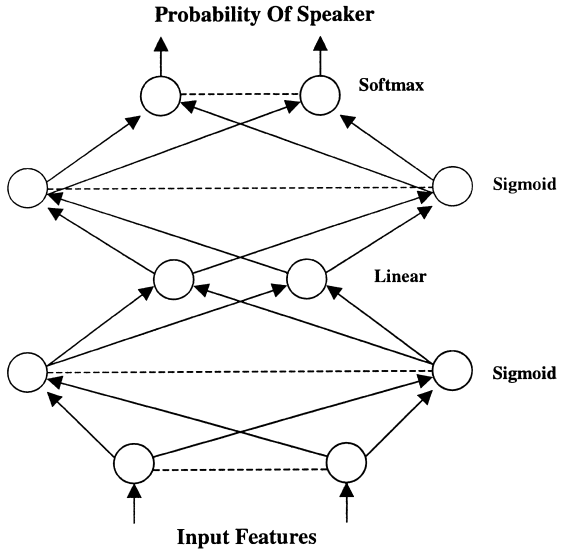

Fig. 2. Example architecture of a 5-layer MLP for speaker recognition. The network can be thought of as the combination of two MLPs, where the MLPs serve the role of feature extraction and classification, respectively. The MLP has a large hidden layer followed by a small "bottleneck" representing the reduction of the feature space. The bottleneck is followed by a large nonlinear hidden layer and a softmax output layer.

3.2. MLP network configuration

The feature mapping function $\mathscr{F}(X,\Psi)$ and the classifier $Y(\mathscr{F}(X,\Psi);\Lambda)$ can be thought of as two separate single-hidden-layer MLP neural networks that are combined to form a single 5-layer MLP with 3-hidden-layers (see Fig. 2). The MLP has a large input layer with the input feature vector X consisting of consecutive frames of cepstral coefficients (or spectral energies), i.e.,

$$X_t = \left[S_{t+k}^{\mathrm{T}} \ldots S_t^{\mathrm{T}} \ldots S_{t-k}^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{2}$$

where S_t is a vector of cepstra (or spectral energies) computed from a windowed frame of data at time t. The first hidden layer of the network uses a large number of nonlinear (sigmoidal) hidden units, a small "bottleneck" second linear hidden layer, a large third nonlinear hidden layer, and a softmax output layer. The first two layers constitute the feature extractor, $\mathscr{F}(X,\Psi)$, which nonlinearly projects the high number of input features to a lower dimensional space. The last two layers function as the closed-set speaker classifier, i.e., $Y(\mathscr{F}(X,\Psi);\Lambda)$.

The motivation for using a 5-layer MLP with a bottleneck is based on the intuition that the first three layers act as the feature extractor, and the last two layers serve the role of classification. If the network is trained with a large number of speakers, then the feature extractor portion of the network can be retained as a general-purpose feature mapper for robust telephone-based speaker recognition (i.e., it is assumed to be speaker independent, having good speaker discrimination power and handset robustness over general speaker populations). The classifier, on the other hand, is specific to the particular speaker population in the development set and is therefore discarded. Fig. 3 illustrates this idea, with the first three layers of the 5-layer network serving the purpose of extracting features that are then passed on to a GMM speaker recognition classifier. When used in subsequent speaker recognition systems, the resulting feature extractor remains fixed, retaining the "memory" of how to compensate for

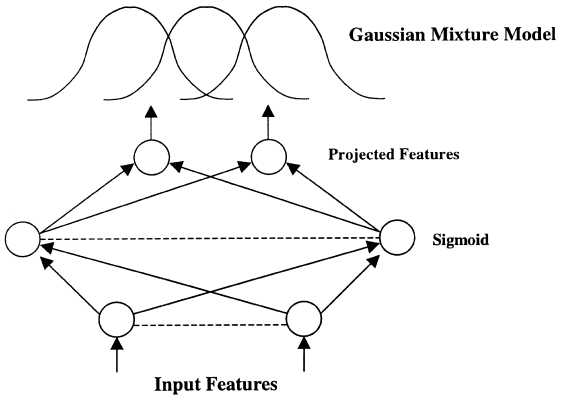

Fig. 3. Architecture of a 3-layer feature transformation MLP for speaker recognition. The MLP was obtained by discarding the last two layers of a 5-layer "bottleneck" MLP trained to maximize speaker recognition performance under mismatched handset conditions.

mismatched telephone handsets in the feature space. This approach is particularly useful in speaker verification applications when the user enrolls on only one phone call (i.e., one type of handset) but uses the system on other handsets as well. This will be demonstrated experimentally in Section 4.

3.3. Initialization of the network

One of the main difficulties in training a 5-layer MLP is that the network is sensitive to the initialization of the weights due to the high number of parameters and hidden layers. If the standard approach of using "small random values" is used to initialize a large 5-layer MLP, the classification performance of the net is often poor (i.e., trapped in a local maximum). The poor performance arises because the weight update values in most cases decay geometrically as the errors are backpropagated from the output layer to the input layer, leading to rates of weight evolution that sometimes differ drastically from layer to layer (Lehr, 1996). This will often cause severe saturation in the input layer. The end result is that either the weights diverge (because of the large learning rate parameter), or the network converges very slowly to an unsatisfactory local minimum.

To overcome the problems with initializing our 5-layer MLP, we use a technique described in (Lehr, 1996). In this approach, initial weights are chosen for each hidden layer such that saturation is largely avoided. The initial weights are independent and identically distributed, and drawn from a uniform zero-mean distribution of variance γ^2/N. The approach sets the initial distribution of weights in a layer receiving N inputs according to

$$w \sim \text{unif}\left[+\sqrt{\frac{3}{N}}\gamma, -\sqrt{\frac{3}{N}}\gamma\right], \tag{3}$$

where γ is a constant between 1 and 1.5 that controls the initial degree of sigmoidal saturation.

4. Experimental results

4.1. Initial feature analysis

Referring to Fig. 1, the initial feature analysis component of the speaker recognition system consisted of the standard SRI mel-cepstral processing component (Murthy et al., 1999) and an estimate of pitch. The mel-cepstral coefficients were computed by applying a sliding 25 ms window to the speech, resulting in a frame of speech every 10 ms. Each frame of speech was transformed to the frequency domain via a 256-point fast Fourier transform (FFT). The frequency scale was warped according to the mel-scale to give a higher resolution at low frequencies and a lower resolution at high frequencies. The frequency scale was multiplied by a bank of 24 filters. The width of each of these filters ranges from the center frequency of the previous filter to the center frequency of the next filter. The filterbank energies were then computed by integrating the energy in each filter, and a discrete cosine transform (DCT) was used to transform the filterbank log-energies into 17 mel-cepstral coefficients. CMS was applied to all frames (Furui, 1981). For the estimation of the pitch, we used an auditory model-based pitch tracker (Weintraub, 1985). The pitch tracker uses a model of cochlear filtering to compute autocorrelation-like functions and dynamic programming for tracking and voiced/unvoiced decisions.

The output of the initial feature analysis was used to construct a large input vector for the 5-layer MLP in the nonlinear feature transformation component of Fig. 1. The 162 inputs for the MLP consisted of the 17 mel-cepstral coefficients and the estimate of the pitch for the current frame, and four past frames and four future frames.

4.2. Results on development database

To train the 5-layer MLP, we used approximately 2 hours (855 sentences) from the 1996 NIST Speaker Recognition corpus (Przybocki and Martin, 1998). The NIST corpus is a subset of Switchboard, a conversational-style corpus of long distance telephone calls. The sentences were selected from a population of 31 speakers (16 male, 15 female), where each speaker was recorded over multiple telephone handsets. The handset labels for the telephone calls were determined by an automatic handset detector that was specifically developed to label the Switchboard corpus (Heck and Weintraub, 1997). The handset detector was implemented as a maximum-likelihood classifier based on a 1024-order GMM. It was trained on the SRI ATIS corpus (Murthy et al., 1999) to discriminate between speech recorded on a telephone handset with a carbon-button microphone and a handset with an electret microphone. A standard mel-cepstra front end was used as the feature set with linear filtering compensation (CMS) applied before training and testing of the handset detector.

To examine the importance of the pitch input, the 9-frame temporal window, and the degradation loss as a result of the dimension reduction from 162 inputs to 34 hidden units in the bottleneck layer, we trained several MLPs and tested their frame-level cross-validation performance on identifying the 31 speakers in the development database. The total number of vectors from the NIST dataset that were used for these tests was 765,060 with 687,156 vectors used for training and 77,904 for cross-validation. The results of these experiments are shown in Table 1. The first column describes the inputs used, the second column describes the network architecture, and the last column shows the resulting frame-level performance on the cross-validation dataset. The MLP that serves as our baseline is denoted as 'MLP5-34' in the second row. It has 500 sigmoidal units in the first hidden layer, a bottleneck layer with 34 linear units, a third hidden layer with 500 sigmoidal units, and a final softmax layer with 31 outputs (one for each speaker in the development set). To determine the impact of the bottleneck, we trained a network without a bottleneck, i.e., a 3-layer MLP (one hidden layer) denoted as 'MLP3'. This MLP has the same number of inputs, 500 hidden units, and 31 outputs. To study the effect of the pitch feature, we trained the MLP named 'MLP5-NO-PITCH', which is the same as the baseline 5-layer network but without pitch information (only 153 inputs). Finally, to test the effect of the large temporal window, we trained the MLP named 'MLP5-1frame', which is the same as the baseline but with only one input frame (as compared to the nine frames used in the other systems).

Several observations can be made from the results shown in Table 1. First, when comparing the first two rows, it can be seen that there is loss in performance due to the bottleneck. Other observations from the table are that pitch and the temporal window both help speaker identification performance. Comparing the second row to the fourth row, the addition of the pitch estimate to the input feature vector yields a 3% absolute gain. Comparing the second and fifth rows, we observe a 10.3% absolute gain from the temporal window.

Table 1
Frame-level results on the cross-validation portion of the development corpus (NIST, 1996 Speaker Recognition Evaluation)

Inputs	Name	Frame correct
9 Frames + pitch	MLP3	37.2%
9 Frames + pitch	MLP5-34	28.9%
9 Frames, no pitch	MLP5-NO-PITCH	25.9%
1 Frame + pitch	MLP5-1frame	18.6%

4.3. Results on evaluation database

The 1998 NIST Speaker Recognition Evaluation corpus (Przybocki and Martin, 1998) is focused on the task of speaker detection, where the task is to determine whether a specified target speaker is speaking during a given speech segment. This task is posed in the context of conversational telephone speech with limited training data.

The test corpus has 500 speakers (250 male and 250 female), with the 500 speakers serving as both target speakers and as nontarget (impostor) speakers. There are three training conditions for each target speaker. Two of these conditions use 2 minutes of training speech data from the target speaker, while the other training condition uses more than 2 minutes of training speech data. The conditions are

- *1-Session training*. 2 minutes of speech data taken from only one conversation.
- *2-Session training*. Equal amounts of training data taken from two different conversations collected from the same phone number.
- *2-Session-full training*. All available speech data taken from two different conversations collected from the same phone number.

The actual duration of the training files used for the 1-Session and 2-Session training conditions was approximately 2 minutes, whereas the duration of the 2-Session-full condition varied from 2 to 5 minutes.

Performance on this corpus was computed and evaluated separately for female and male target speakers and for the three training conditions. For each of these training conditions, there are three different test conditions of interest:

- *Test segment duration*. Performance was computed separately for three different test durations. These durations were nominally 3, 10, and 30 s.
- *Same/different phone number*. Performance was computed separately for test segments from the training phone number versus those segments from different phone numbers. The handset type label (electret or carbon-button) was the same as that used in training.
- *Same/different handset type*. Performance was computed separately for test segments with the same handset type label as training, versus segments with a different handset label. All test segments were from phone numbers different from the training number.

A total of nine tests constitute the NIST evaluation: one for each of the three test durations and for each of the three training conditions. There is an average of 10 test segments for each target speaker over the nine tests, totaling approximately 45,000 target speaker trials and about 405,000 nontarget speaker trials.

The formal evaluation measure used in the NIST evaluation was a detection cost function (DCF), defined as a weighted sum of the miss and false alarm error probabilities:

$$\mathrm{DCF} = C(\mathrm{miss})P(C)P(\mathrm{miss}) + C(\mathrm{fa})P(I)P(\mathrm{fa}), \tag{4}$$

where $C(\mathrm{miss})$ and $C(\mathrm{fa})$ are the costs of missing a claimant speaker and falsely accepting an impostor, respectively, $P(C)$ and $P(I)$ are the a priori probabilities of a claimant speaker and an impostor speaker, and $P(\mathrm{miss})$ and $P(\mathrm{fa})$ are the probabilities of missing a claimant and falsely accepting an impostor. The value of $P(I)$ was 0.01, $C(\mathrm{miss})$ was 10, and $C(\mathrm{fa})$ was 1.

For comparison, we implemented a state-of-the-art baseline system using Bayesian-adapted GMMs and a standard mel-cepstral front end (Reynolds, 1997b). Concatenated mel-cepstra, Δ-cepstra and $\Delta\Delta$-cepstra with the corresponding energy terms (E, ΔE and $\Delta\Delta$E) are used as acoustic observations in the experiments. CMS is used for channel equalization in all experiments. The classifier of the baseline system is a GMM

$$p(x_t \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\boldsymbol{x}), \tag{5}$$

with mixture weights p_i and Gaussian densities $b_i(\boldsymbol{x})$. The GMM for the target speaker is created by adapting a large speaker-independent GMM representing the general (impostor) population of the same gender as the target speaker. The impostor model is also used to normalize the score of the target speaker, where the score of the target speaker is computed as the average log-likelihood of the utterance $X = \{x_1, \ldots, x_T\}$,

$$\mathscr{L}(X \mid \lambda) = \frac{1}{T}\sum_{t=1}^{\mathrm{T}} \log p(x_t \mid \lambda), \tag{6}$$

and the normalization of the score with the impostor model is implemented as a log-likelihood difference,

$$\Lambda(X \mid s) = \mathscr{L}(X \mid \lambda_s) - \mathscr{L}(X \mid \lambda_I), \tag{7}$$

with Λ_s and Λ_I denoting the target and impostor speaker model scores, respectively.

To improve the robustness of the baseline system, the impostor model is trained with speakers that use the same telephone handset type as that used by the target speaker during the enrollment session (Heck and Weintraub, 1997). As described in Section 1, this approach gave a 60% improvement in performance (as compared to a general handset-independent impostor model). Only two handset types were assumed to be used in the NIST corpus: electret and carbon-button. With the two genders and two handset types, we built four separate impostor models for score normalization.

To construct a system with the new MLP-based features, we discarded the last two layers of the 5-layer MLP trained on the development database and fixed the weights of the first three layers to serve as a general-purpose feature transformation. We used this 3-layer MLP as an additional transformation on the mel-cepstral features described above. The classifier was the same as described above: a Bayesian-adapted GMM system with speaker-independent, gender-dependent, and handset-dependent impostor models for score normalization. Tests were completed with this new feature extractor (initial feature analysis followed by the nonlinear feature transformation) on the 1998 NIST Speaker Recognition Evaluation corpus.

Tables 2 and 3 show the performance of the new MLP-based feature transformation technique of this paper for the 1998 NIST Speaker Recognition Evaluation corpus. Table 2 shows equal error rates (EER), where the equal error rate is defined as the percentage of errors observed when the threshold is set a posteriori to yield equal false accept and false reject rates. Table 3 shows the DCF as defined in Eq. (4). Male and female results are shown separately, with all training and testing conditions displayed in the columns. The rows are distinguished by the type of feature used by the system, with the first row labeled "cepstrum" denoting a standard mel-cepstra feature set with CMS and handset-dependent normalizing impostor models (described as the baseline system earlier in this section). The second row labeled "MLP5-34" are the new features developed in this paper. The third row labeled "cepstrum(hnorm)" is a system with the standard mel-cepstra feature set but with the

Table 2
Equal error rate (EER) performance (in percent) on the 1998 NIST Speaker Recognition Evaluation corpus for multiple training conditions and test durations

Type of feature	1-Session			2-Session			2-Session-full		
	3 s	10 s	30 s	3 s	10 s	30 s	3 s	10 s	30 s
Male									
Cepstrum	22.7	18.6	17.3	22.0	17.8	16.5	21.2	17.2	16.5
MLP5-34	23.6	16.9	14.2	22.1	15.5	12.8	19.8	14.6	11.9
Cepstrum(hnorm)	20.3	14.9	12.9	18.9	14.0	11.8	18.4	13.2	11.6
MLP5-34 + cepstrum(hnorm)	18.8	13.6	11.4	17.6	12.5	10.2	16.5	11.8	9.8
Female									
Cepstrum	23.3	17.6	16.4	21.7	16.9	15.8	21.1	16.8	15.5
MLP5-34	24.5	16.0	13.7	23.2	14.8	12.5	20.7	14.0	11.5
Cepstrum(hnorm)	20.2	14.8	12.1	19.2	13.2	11.0	18.5	13.3	11.0
MLP5-34 + cepstrum(hnorm)	18.8	12.7	10.8	17.2	11.4	9.4	16.6	11.2	9.8

Table 3
Detection cost function (DCF) performance ($\times 10^3$) on the 1998 NIST Speaker Recognition Evaluation corpus for multiple training conditions and test durations

Type of feature	1-Session			2-Session			2-Session-full		
	3 s	10 s	30 s	3 s	10 s	30 s	3 s	10 s	30 s
Male									
Cepstrum	78.8	65.0	61.0	76.4	62.1	58.4	72.9	58.3	55.2
MLP5-34	83.8	67.4	56.4	77.0	60.7	52.3	73.7	57.9	56.4
Cepstrum(hnorm)	78.6	65.5	53.8	74.6	61.3	49.8	72.2	60.4	48.6
MLP5-34 + cepstrum(hnorm)	72.1	56.4	46.0	67.1	52.5	43.1	63.3	51.1	46.0
Female									
Cepstrum	84.7	70.6	65.8	82.6	68.5	64.7	79.8	66.2	63.6
MLP5-34	83.7	65.2	54.0	82.4	60.6	49.3	77.6	56.8	46.2
Cepstrum(hnorm)	81.3	62.0	54.6	79.2	57.2	50.0	75.9	57.6	50.0
MLP5-34 + cepstrum(hnorm)	72.2	52.4	44.3	69.4	47.4	40.0	66.6	47.4	40.2

scores postprocessed by the HNORM technique developed recently by Reynolds (1997b) (described in Section 1). Finally, the fourth row labeled "MLP5-34 + cepstrum(hnorm)" is the combination of the systems from the second and third rows. These systems were combined by a weighted average of their output scores. The weights were 0.7 for the "cepstrum(hnorm)" system, and 0.3 for the "MLP5-34" system except for the 3 s cases, where the weights were 0.6 and 0.4, respectively. These weights were selected a priori through an optimization completed on the development database.

Several observations can be made from the results shown in Table 2. Comparing the MLP-based features developed in this paper with the baseline cepstrum system using CMS, the MLP-based system shows an EER reduction of 15–28% (relative) for the longer test utterances. The EER results are mixed for the short 3 s tests except for the 2-Session-Full training condition, suggesting that a longer training session compensates for the shorter test duration with the MLP features. Comparing the third and fourth rows (for each gender), the combination of the "MLP5-34" with the "cepstrum(hnorm)" systems yields approximately 15% improvement. Comparing the baseline "cepstrum" system in the first row with the combination of the MLP and cepstrum(hnorm) in the fourth row, we observe a 21–40% improvement in EER.

Table 3 shows similar performance gains for the DCF. Comparing the first two rows, the MLP-based features show a 27% improvement in DCF for the females, but mixed results for the males. Comparing the cepstrum(hnorm) and combined MLP5-34 + cepstrum(hnorm), the MLP-based features improve the performance by 5–15% for males, and 12–20% for females. As with EER, the largest improvements are observed when combining the MLP-based features with the cepstrum using HNORM, giving between 7% and 38% improvement over the baseline cepstrum system.

Table 4 shows the performance of the MLP features with respect to the training–testing handset combinations. EER and DCF values are shown for combined male and female tests on the 1-Session training condition and 30 s test length. The notation "X–Y" refers to the X handset type used during training, and the Y handset type used during testing (where E is electret, and C is carbon-button). As can be seen from the first two rows of the table, we get mixed results on matched handset conditions, but consistent improvements for the mismatched handsets. The largest improvement is with the "E–C" condition, i.e., training on electret and testing on carbon-button handsets. Comparing the first and last rows of the table, we get better results with all combinations except for the DCF on "C–C" (carbon–carbon).

Table 4
Performance with respect to training–testing handset combinations[a]

Type of feature	E–E		C–C		E–C		C–E	
	EER	DCF	EER	DCF	EER	DCF	EER	DCF
Cepstrum	11.6	60.6	24.7	47.1	33.9	98.9	26.2	96.9
MLP5-34	14.0	64.2	17.9	64.9	23.0	90.5	19.6	84.4
Cepstrum(hnorm)	11.3	58.0	19.7	63.6	23.5	82.6	23.5	82.6
MLP5-34 + cepstrum(hnorm)	9.3	50.4	17.4	56.4	21.2	78.2	14.4	69.9

[a] Equal error rates and detection cost function values are shown for combined male and female tests on the 1-Session training condition and 30 s test length. The notation "X–Y" refers to the X handset type used during training, and the Y handset type used during testing (where E is electret, and C is carbon-button).

Table 5
Correlation coefficients between MLP5-34 and cepstrum speaker recognition systems on NIST evaluation

Test length	Male	Female
3	0.61	0.47
10	0.68	0.71
30	0.76	0.77

Finally, Table 5 shows the correlation coefficients between the MLP5-34 and the cepstral-based speaker recognition system. These values were computed on the log-likelihood ratio scores for the three different test lengths on the 1-Session training condition for the NIST evaluation corpus. The low values indicate that the MLP5-34 is providing additional information not seen in the cepstral feature stream that can be utilized for improved speaker recognition performance. This is supported by the performance gains seen in the earlier speaker detection results.

5. Conclusions

A discriminative feature design technique produces speaker recognition features robust to telephone handset distortions. Our results on the 1998 NIST Speaker Recognition Evaluation show improvements as high as 28% for the new MLP-based features as compared to a standard mel-cepstral feature set with CMS and handset-dependent normalizing impostor models. If a system is constructed using only the new MLP-based features, the new features should be used for test utterances longer than 3 s, and for mismatched handset conditions. On the other hand, if the system uses a combination of the MLP-based features and the cepstral features with HNORM, then the system should be used for all test lengths and handset combinations. The MLP-based feature design approach of this paper can be extended to other types of input data such as speech over cellular phones and speaker-phone speech. In addition, a wider range of input representations and resolutions can be utilized with this approach such as first and second derivatives of cepstrum, filter-bank energy levels, and different analysis windows. Finally, we note that although the training of the MLP with five layers is computationally expensive ($25 \times$ real time), the application of the MLP3 in a feed-forward mode is very fast (less than $0.4 \times$ real time). Thus the approach is feasible in realistic settings.

Acknowledgements

The authors thank Francoise Beaufays for providing significant technical advice related to this work and making suggestions for improving this manuscript. In addition, the authors thank Mark Przybocki and Alvin Martin at NIST for providing the scoring code for the experiments presented in this paper.

References

Baum, E.B., Wilczek, F., 1988. Supervised learning of probability distributions by neural networks. In: D. Anderson, (Ed.), Neural Information Processing Systems. pp. 52–61.

Bengio, Y., De Mori, R., Flammia, G., Kompe, R., 1992. Global optmization of a neural network-hidden markov model hybrid. IEEE Trans. Neural Networks 3 (2).

Chengalvarayan, R., Deng, L., 1997. HMM-based speech recognition using state-dependent, discriminatively derived transforms on mel-warped DFT features. IEEE Trans. Speech and Audio Process. 5 (3), 243–256.

Euler, S., 1995. Integrated optimization of feature transformation for speech recognition. In: Proceedings European Conf. on Speech Communication and Technology. EUROSPEECH. pp. 109–112.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. ASSP-29, 254–272.

Heck, L.P., Weintraub, M., 1997. Handset dependent background models for robust text-independent speaker recognition. In: Proceedings Internat. Conf. on Acoust. Speech and Signal Process.

Hermansky, H., Morgan, N., Bayya, A., Kohn, P., 1991. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). In: Proceedings European Conf. on Speech Communication and Technology. EUROSPEECH. pp. 1367–1370.

Lehr, M., 1996. Scaled stochastic methods for training neural networks, Ph.D. Dissertation, Information Systems Laboratory, Dept. of Electrical Engineering, Stanford University.

Liu, F.-H., Stern, R.M., Acero, A., Moreno, P.J., 1994. Environment normalization for robust speech recognition using direct cepstral comparison. In: Proceedings Internat. Conf. on Acoust. Speech Signal Process., Vol. 2, pp. 19–22.

Mammone, R.J., Shang, X., Ramachandran, R.P., 1996. Robust speaker recognition. IEEE Signal Process. Magazine 13, 58–71.

Murthy, H.A., Beaufays, F., Heck, L.P., Weintraub, M., 1999. Robust text-independent speaker identification over telephone channels. IEEE Trans. Speech and Audio Process. 7, 554–568.

Neumeyer, L., Weintraub, M., 1994. Probabilistic optimum filtering for robust speech recognition. In: Proceedings Internat. Conf. on Acoust. Speech Signal Process., pp. 417–420.

NIST, 1996. Speaker recognition workshop. In: NIST Workshop Notebook, Linthicum Heights, Maryland.

NIST, 1997. Speaker recognition workshop. In: NIST Workshop Notebook, Linthicum Heights, Maryland.

NIST, 1998. Speaker recognition workshop. In: NIST Workshop Notebook, Linthicum Heights, Maryland.

Paliwal, K.K., Bacchiani, M., Sagisaka, Y., 1995. Minimum classification error training algorithm for feature extractor and pattern classifier in speech recognition. In: Proceedings European Conf. on Speech Communication and Technology. EUROSPEECH. pp. 541–544.

Przybocki M.A., Martin, A.F., 1998. NIST speaker recognition evaluations. In: LREC, Granada, Spain, pp. 331–335.

Quatieri, T.F., Reynolds, D.A., O'Leary, G.C., 1998. Magnitude-only estimation of handset nonlinearity with application to speaker recognition. In: Proceedings Internat. Conf. on Acoust. Speech Signal Process., Vol. 2, pp. 745–748.

Rahim, M., Bengio, Y., Lecun, Y., 1997. Discriminative feature and model design for automatic speech recognition. In: Proceedings European Conf. on Speech Communication and Technology. EUROSPEECH. Rhodes, Greece.

Reynolds, D.A., 1995. Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17, 91–108.

Reynolds, D.A., 1997a. Htimit and llhdb: Speech corpora for the study of handset transducer effect. In: Proceedings Internat. Conf. on Acoust. Speech Signal Process., Vol. 2, pp. 1535–1538.

Reynolds, D.A., 1997b. Comparison of background normalization methods for text-independent speaker verification. In: Proceedings European Conf. on Speech Communication and Technology. EUROSPEECH.

Richard, M.D., Lippmann, R.P., 1991. Neural network classifiers estimate bayesian a posteriori probabilities. Neural Computation 3 (4), 461–483.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Parallel Distributed Processing. MIT Press, Cambridge, pp. 318–364.

Stern, R.M., Liu, F.-H., Moreno, P.J., Acero, A., 1994. Signal processing for robust speech recognition. In: Proceedings Internat. Conf. on Spoken Language Process., Vol. 3, pp. 1027–1030.

Weintraub, M., 1985. A theory and computational model of auditory monaural sound separation, Ph.D. Dissertation, Stanford University.