IOSR Journal of Electronics and Communication Engineering (IOSR-JECE)
e-ISSN: 2278-2834,p- ISSN: 2278-8735.Volume 9, Issue 5, Ver. II (Sep - Oct. 2014), PP 31-36
www.iosrjournals.org

Optimization of Telephone Handset Distortion in Speaker Recognition by Discriminative Feature Design

Kadiri Kamoru Oluwatoyin
Electrical and Electronic Engineering Department, Federal Polytechnic Offa, Kwara State. Nigeria

Abstract: *This technique used mel-cepstral features, log spectrum and prosody based features with a non-linear artificial neural network in designing speaker recognition features that minimize telephone handset distortion. Effect of handset distortion was done to maximize speaker recognition performance specifically in the setting of telephone handset mismatch between training and testing as results on the 1998 NIST Speaker Recognition Evaluation corpus shows a high efficiency of 98% performance.*

Keywords: *Speaker; Channel; Telephone handset; distortion; mel-cepstral; log spectrum; Neural network; design.*

I. Introduction

Tasks that are easily performed by humans, such as face or speech or speaker recognition, prove difficult to perform with computers. Speaker recognition involves two tasks: identification and `verification. In identification, the goal is to determine which voice in a known group of voices best matches the speaker. In verification, the goal is to determine if the speaker is who he or she claims to be as some factors such as background of the speaker, the handset or microphone use, etc. affect the performance of speaker recognition system.

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: model-based, score-based, and feature-based.

i. **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.

ii. **A score-based handset and channel compensation method** for speaker recognition systems called HNORM was presented in Reynolds, 1997b. The method utilized an automatic handset detector during the training of the speaker model. 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. 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.

iii. **Feature-based methods:** 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.

II. Aim And Objectives

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).

III. Research Problems

I. Our approach specifically focuses on the problem of telephone handset mismatch between training and testing
II. 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 became more robust to channel distortions.

IV. Statement of The Problems

There is the need for the development of a discriminative feature design approach for speaker recognition because the linear mapping needs to be initialized to produce standard cepstral-based features, Output feature vectors from the nonlinear neural network needs to be added to the linear cepstral feature vectors, and the resulting single modified feature vector needs to be fed into the HMM classifier.

V. Literature Review

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 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 Murthy et al. (1999), a new filter bank design was introduced 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 from 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 features design techniques have been studied mainly for the speech recognition task Bengio (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 (1992). Euler (1995) reports 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 (1997).

VI. Methodology

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, being bandlimited and 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 described 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 as followed.
The Algorithms
The feature vectors for each speaker X is given by
$X_i = [X_1\ X_2\ X_3\ \ldots\ldots\ X_t]$ which is the sequence feature vectors belonging to speaker i

i = 1, 2, 3 … N which is the total number of speakers
The set of speakers is given by
$S = [S_1\ S_2\ S_3\ \ldots.\ S_N]$

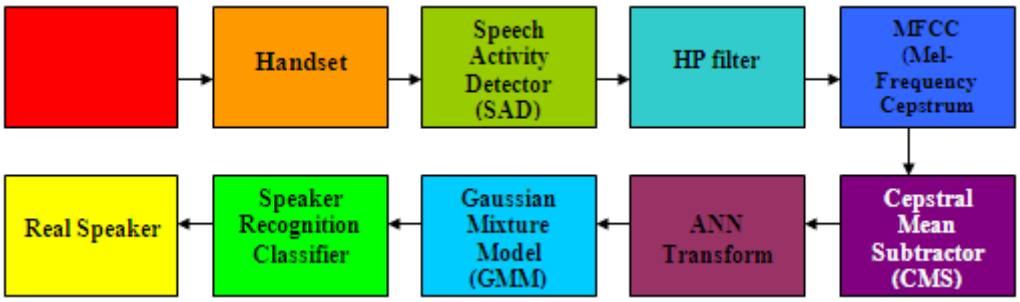

Fig 1 Block diagram of the discriminative feature design

The objective function J which is the cross entropy cost functions so we minimize J so that we can maximize the speaker recognition performance.

Minimize $\left[\left[J = -E X\{ \sum [di \log Yi(f(X,\Psi);\Lambda)] + (1-d)\log(1-Yi(f(X,\Psi);\Lambda)\ \} \right]\right]$ .....1 the constrains are

E = expectation over the dataset
f (X,Ψ) = mapping of input fature X with the corresponding set of parameters Λ and Ψ of the parameter of the classifier
Yi f[(X,Ψ); Λ] = is the ith output of the speaker recognition
di = desired speaker decision

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. Secondly, it maximizes a posteriori probabilities of the speakers which lead 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).

The initial feature analysis component of the speaker recognition system consisted of the standard SRI mel-cepstral processing component hand 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 filter bank energies were then computed by integrating the energy in each filter, and a discrete cosine transform (DCT) was used to transform the filter bank log-energies into 17 mel-cepstral coefficients. CMS was applied to all frames. For the estimation of the pitch, we used an auditory model-based pitch tracker. The pitch tracker uses a model of cochlear filtering to compute autocorrelation- like functions and dynamic programming for tracking and voiced/unvoiced decisions.

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:

DCF = C(miss) P(C)P(miss) + C(false)P(I)P(false)
Where C.(miss) = the costs of missing a claimant speaker
C(false) = the cost of falsely accepting an impostor
P(C) = the a priori probabilities of a claimant speaker
P(I) = priori probability of a impostor speaker
P(miss) = the probabilities of missing a claimant
P(false) = the probability of falsely accepting an impostor through schematic diagram for speaker

Verification Section

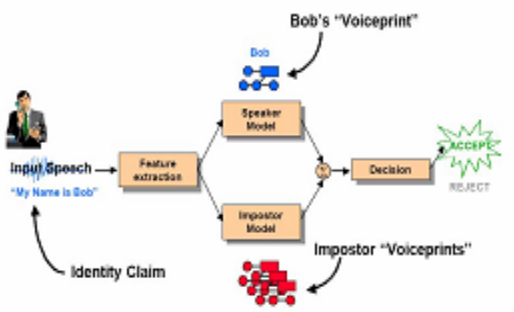

Figure 2: Schematic diagram of speaker verification

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, D-cepstra and DD-cepstra with the corresponding energy terms (E, DE and DDE) 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

$$P(x|\lambda) = \sum_{i=1}^{N} pibi(x)$$

Where

P_i = mixture weight

bi(x) = Gaussian densities

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

$X = \{X_1\ X_2\ X_3\ \ldots\ X_N\}$

$$L(X|\lambda) = \frac{1}{N}\sum_{i=1}^{N}\log p(x_i \mid \lambda)$$

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

$$\Lambda(X \mid s) = L(X \mid \lambda_s) - L(X \mid \lambda_i)$$

Where

Λ_s and Λ_I = the target and impostor speaker model scores, respectively

VII. Data Analysis

The number of languages currently estimated and catalogued in Nigeria is **521**. This number includes 510 living languages, two second languages without native speakers and 9 extinct languages. In some areas of Nigeria, ethnic groups speak more than one language. The official language of Nigeria is English. The major languages spoken in Nigeria are Hausa, Igbo, Yoruba, Edo, Efik, Adamawa, Fulfulde, Idoma, and Central Kanuri. Even though most ethnic groups prefer to communicate in their own languages, English, being the official language, is widely used for education, business transactions and for official purposes.

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 69 speakers (45 male, 24 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.

There are three training conditions for each captured speaker. Two of these conditions use 2 minutes of training speech data from the captured speaker, while the other training condition uses more than 2 minutes of training speech data.

Dialect Regions	Total Session per language	# male	# female	Training session	Testing session
Yoruba	65	20	10	40	25
Hausa	20	10	6	16	4
Igbo	40	15	8	35	15
Total	125	45	24	91	44

Table 1: composition of the NIST data set

# PHONE	# SPEAKERS per language			**#SESSION per PHONE**
	Yoruba	Igbo	Hausa	
A	10	8	6	**50**
B	10	8	5	**50**
C	10	7	5	**25**

Table 2: No of distinct phone set per person per session

VIII. Summary

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.
The largest improvement is with the ``E-C'' condition, i.e., training on electret and testing on carbon-button handsets.

A discriminative feature design technique produces speaker recognition features robust to telephone handset distortions. A new feature is used for test utterances longer than 3s, and for mismatched handset conditions and a combination of the MLP-based features and the cepstral features with HNORM, then the system is 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, filterbank energy levels, and different analysis windows.

IX. Conclusion

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. 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 captured speaker during the enrollment session. This approach gave a 60% improvement in performance (as compared to a general handset-independent impostor model). 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. ``MLP5-34'' with the ``cepstrum (HNORM)'' systems yields approximately 15% improvement. 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.

Reference

[1]. A.L. Higgins, L.G. Bahler, andJ.E. Porter, "Voice Idenrification Using Nearest- eighbor Distance Measure," Proc. Int. Con! on Acoustics, Speech, and Signal Processing, Minneapolis, 27-30 Apr. 1993, p. 11-375.

[2]. D.A. Reynolds, "Large Population Speaker Recognition Using Wideband and Telephone Speech," SPIE2277, 11 (1994).

[3]. D.A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models," Speech Commun. 17, 91 (Aug. 1995).

[4]. D.A. Reynolds, M.A. Zissman, T.E Quatieri, G.c. O'Leary, and B.A Carlson, "The Effects of Telephone Transmission Degradations on Speaket Recognition Performance," Proc. Int. Con! on Acoustics, Speech, and Signal Processing 1, Detroit, 912 May 1995, p. 329.

[5]. H. Gish and M. Schmidt, "Text-Independenr Speaker Idenrification," IEEE Signal Process. Mag. 11,8 (Oct. 1994).

[6]. H.-S. Liou and R. Mammone, "A Subword Neural Tree Network Approach ro Text-Dependenr Speaker Verification," Proc. Int. Con! on Acoustics, Speech, and Signal Processing 1, Detroit, 9-12 May 1995, p. 357.

[7]. J.J. Godfrey, E.C. Holliman, and J. MacDaniel, "Switchboard: Telephone Speech Corpus for Research and Developmenr," Proc. Int. Con! on Acoustics, Speech, and Signal Processing 1, San Francisco, 23-26Mar. 1992, p. 1-517.

[8]. J.-L. Floch, C. Monracie, and M.-J. Carary, "Investigations on Speaker Characterization from Orphee System Technics," Proc. Int. Con! on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 19-22 Apr. 1994, p. 1-149.

[9]. J.P. Campbell, Jr., "Testing with the YOHO CD-ROM Voice Verification Corpus," Proc. Int. Con! on Acoustics, Speech, and Signal Processing, Detroit 1,9-12 May 1995, p. 341.

[10]. L. Gillick, J. Baker, J. Baker, J. Bridle, M. Hunr, Y. lro, S. Lowe, J. Orloff, B. Peskin, R. Roth, and E Scallone, "Application of Large Vocabulary Conrinuous Speech Recognition ro Topic and Speaker Idenrification UsingTelephone Speech," Proc. Int. Con! on Acoustics, Speech, and Signal Processing, Minneapolis, 27-30 Apr. 1993, p. 11-471.
[11]. L.G. Bahler, J.E. Porter, and A.L. Higgins, "Improved Voice Identification Using a Nearest-Neighbor Distance Measure," Proc. Int. Con! on Acoustics, Speech, and Signal Processing, Adelaide, Australia, 19-22 Apr. 1994, p. 1-321.