Toward 2003 NIST Speaker Recognition Evaluation: The WCL-1 System

Todor Ganchev, Nikos Fakotakis, George Kokkinakis

Wire Communications Laboratory, University of Patras, 26500 Rio-Patras, Greece

tganchev@wcl.ee.upatras.gr

Abstract. A detailed description of our text-independent speaker verification (SV) system, referred to as WCL-1, a participant in the one-speaker detection task of the 2003 NIST Speaker Recognition Evaluation (SRE) is presented. It is an improved version of our baseline system, which has successfully participated in the 2002 NIST SRE. In addition to the short-term spectrum represented by the Mel-frequency scaled cepstral coefficients (MFCCs), the improved WCL-1 system exploits also prosodic information to account for the speaking style of the users. A logarithm of the energy, computed for the corresponding speech frame, replaces the first MFCC coefficient, which was found very much influenced by the transmission channel and the handset characteristics. Furthermore, a logarithm of the fundamental frequency f0 is added to the other parameters, to form the final feature vector. Instead of the traditional ln(f0), we propose $\ln(f0\text{-}f0_{min})$, which we found out to be much more effective, due to its extended dynamic range that better corresponds to the relative importance of the fundamental frequency. The constant $f0_{min}$ is derived as 90% of the minimal fundamental frequency the pitch estimator can detect. Comparative results between the improved WCL-1 system and the baseline version, obtained in the one-speaker detection task over the 2001 NIST SRE database, are reported.

INTRODUCTION

In general, the task of speaker verification is defined as: making a decision, if the identity of a given speaker coincides with the one he/she claims. For that reason, any automatic SV system receives two inputs: an identity claim (PIN, User's name, etc.) made by the speaking person and a certain amount of speech, representing his/her voice. Another frequently used term for description of the SV task is *one-speaker detection*, since the output of the process is always a binary decision: 'yes' – the speaker is accepted with the claimed personality, or 'no' – he is rejected as an impostor. In case of text-independent SV, which is the most challenging among all SV tasks, the speaker is not obligated to follow a specific predefined scenario, as pronouncing a password or prompted by the system sequence of numbers or sentences. Therefore, the SV decision is based only on the user's voice, and not on a specific knowledge he has. That scenario is the most comfortable for the user, because the process of SV remains hidden for him, and the system makes its decision only by using the spontaneous speech collected during the dialog.

Our text-independent SV system is based on Probabilistic Neural Networks (PNN), which were chosen because of their good generalization properties and more importantly because of their fast designing times. PNN design is straightforward and does not depend on training [1]. As a result, PNN are built only for a fraction of the back propagation artificial neural networks training time. It is well known that the PNNs need more neurons compared to back propagation networks, which leads to an increased complexity and higher computational and memory requirements in the process of exploitation. Nevertheless, the SV system described here is capable of working in real-time on common personal computers.

THE WCL-1 SPEAKER VERIFICATION SYSTEM

A simplified block diagram of the WCL-1 system is presented in Fig.1. The lower part of the figure depicts the operational configuration of the system, while the upper part summarizes the process of training. In the training phase, we demonstrate the way that the Universal Background Codebook (UBgCB) is constructed, as well as the building of the personal codebooks and the individual PNNs for the enrolled users. The lower part of the figure reveals the consecutive processing stages of every test trial: from feature extraction, model matching and probability estimation, to making a final decision. The identity claims for all test trials are provided by a test-control file.

In the following paragraphs, for thoroughness of our exposition, a description of the main building blocks of our speaker verification system is presented.

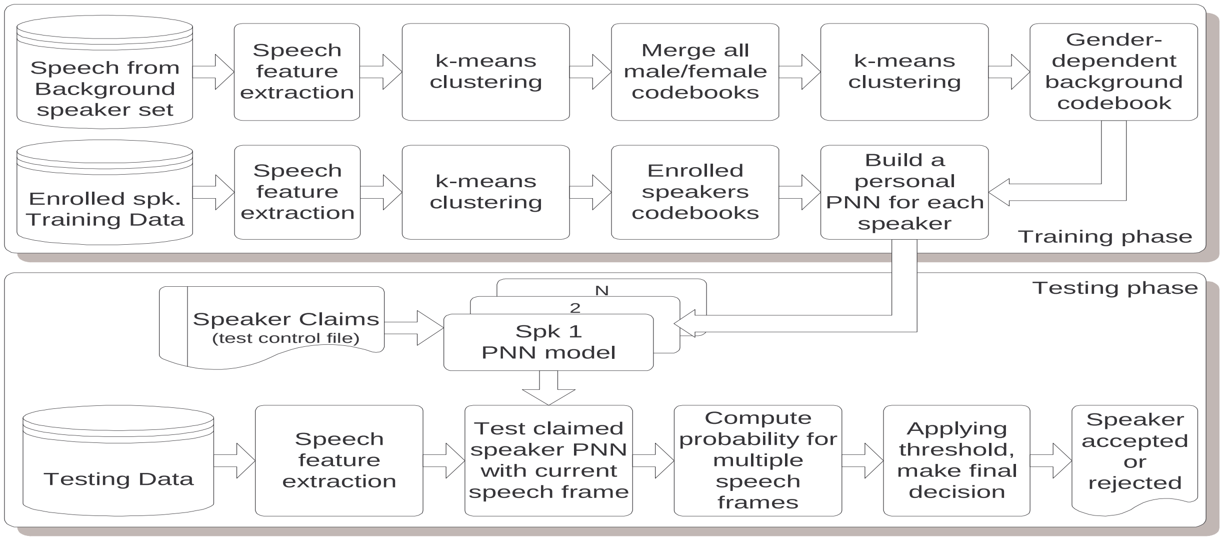

FIGURE 1. A simplified block diagram of the WCL-1 system

Pre-processing and feature extraction

Saturation by level is a common phenomenon for telephone speech signals. In order to reduce the spectral distortions it causes, a band-pass filtering of speech is performed as a first step of the feature extraction process. A fifth-order Butterworth filter with pass-band from 80 Hz to 3800 Hz is used for both training and testing. Then the speech signal, sampled at 8 kHz, is pre-emphasized by the filter $H(z)=1-0.97z^{-1}$ and subsequently, windowed into frames of 40 ms duration, at a frame rate of 100 Hz using a Hamming window. The speech frames formed in that manner are subject to 2048-point Short Time Fourier Transform (STFT). Each frame processed so far has passed through a set of 32 triangular band-pass filter-bank channels. We have accepted an approximation of the Mel-scale, with 13 linearly spaced filter-banks, lowest central frequency 200 Hz, highest 1000 Hz and 19 log-spaced (factor 1.0711703) with highest central frequency 3690 Hz. Finally, 32 dimensional feature vectors are formed, after applying Discrete Cosine Transform to the log-filter-bank outputs. Only the feature vectors extracted for voiced speech frames are used to represent the speakers' identity. The voiced/unvoiced speech separation is performed by the modified autocorrelation method with clipping [2].

In this year's version of the WCL-1 system, the logarithm of the energy, computed for the corresponding speech frame, replaces the first MFCC coefficient. Furthermore, the logarithm of the fundamental frequency f0 is added, to form the final feature vector. In fact, instead of the traditional ln(f0), we propose $\ln(f0-f0_{min})$, which we found out to be much more effective, due to the extended dynamic range that better corresponds to the relative importance of the fundamental frequency. The constant $f0_{min}$= 55 Hz is selected, as 90% of the minimal fundamental frequency the pitch extractor can detect -- in our case 60Hz.

Construction of the Codebooks

The feature vectors extracted from the voiced speech frames are used to model the speaker's voice. Because the complexity of the PNNs depends strongly on the number and dimensionality of the training vectors, the k-means clustering algorithm [3] is used to reduce their amount. A pre-initialization of the initial cluster centres, by running k-means over a smaller dataset, consisting of about 10% of the available data, is performed. Codebooks are built, both for the enrolled users and for the world of possible impostors. It was experimentally found, that a codebook composed of 128 vectors is large enough to maintain a good representation accuracy of the speakers' identity, and one consisted of 256 vectors leads to a slightly better performance. In this year's WCL-1 system, we make use of 256-vectors sized codebook of in order to maximize the quality of the target users' models, in contrast to the baseline system, where 128 vectors were used. For the background reference models, a codebook of at least 256 vectors is necessary, and for achieving a better accuracy 512, 1024, or 2048, vectors are recommended.

In the development experiments, performed during the preparation for the 2003 NIST SRE, the target users were enrolled from the 2001 NIST SRE database, and the 2002 NIST SRE database was used for impostor modelling.

The gender-dependent UBgCBs were built by using all the available in the 2002 NIST SRE training speech. In total, 139 male and 191 female speakers were available, each one having about two minutes of speech. As shown in Fig.1, first a personal codebook, consisting of 256-vectors, for each background speaker is created. Then, these background codebooks are merged by gender, and a separate UBgCB is constructed for the male and female speakers. On the next step, the UBgCB size is reduced, by using the k-means clustering technique. In this year's system, an UBgCB consisting of 1024 vectors is utilized, instead of the one composed of 256 vectors which is exploited in the baseline system.

The UBgCB along with the codebooks built for the enrolled speakers are then employed, to design an independent PNN for each target user.

The PNN description

The PNNs implement estimator by using a mixture of Gaussian basis functions (see [1] for details). If a PNN for classification in *K* classes is considered, the probability density function $f_i(\boldsymbol{x}_p)$ of each class κ_i is defined by (1),

$$\mathrm{f}_i(\boldsymbol{x}_p) = \frac{1}{(2\pi)^{d/2}\sigma^d}\frac{1}{M_i}\sum_{j=1}^{M_i}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{x}_p-\boldsymbol{x}_{ij})^T(\boldsymbol{x}_p-\boldsymbol{x}_{ij})\right), \quad \text{with } i=1,\dots,K \tag{1}$$

where $\boldsymbol{x}_{ij}$ is the *j*-th training vector from class κ_i, $\boldsymbol{x}_p$ is the *p*-th input vector, d is the dimension of the speech feature vectors, and M_i is the number of training patterns in class κ_i. Each training vector $\boldsymbol{x}_{ij}$ is assumed a centre of a kernel function, and consequently the number of pattern units in the first hidden layer of the neural network is given as a sum of the pattern units for all the classes. The variance c acts as a smoothing factor, which softens the surface defined by the multiple Gaussian functions. As seen in (1), c has the same value for all the pattern units, therefore, a homoscedastic PNN is considered.

After the probability density function (1) of each class, for a given input vector, is computed, the Bayesian decision rule (2) is applied to distinguish class κ_i, to which the input vector $\boldsymbol{x}_p$ belongs:

$$D(\boldsymbol{x}_p) = \operatorname*{argmax}_i\{h_i c_i \mathrm{f}_i(\boldsymbol{x}_p)\}, \quad i=1,\dots,K \tag{2}$$

where h_i is a-priori probability of occurrence of the patterns of category κ_i, and c_i is the cost function in case of misclassification of a vector belonging to class κ_i. The averaged for all test vectors $\boldsymbol{X} = \{\boldsymbol{x}_p\}, p=1,\dots,P$ probability, for a given test trial to belong to class κ_i is computed by (3):

$$\mathrm{P}(k_i \mid \boldsymbol{X}) = \frac{1}{P}\sum_{p=1}^{P}\boldsymbol{D}(\boldsymbol{x}_p) \tag{3}$$

where $\boldsymbol{X}$ is a *P* x *d* matrix, containing *P* feature vectors, each one with dimensionality *d*. For every trial, the averaged probability for all output decisions of a particular PNN, obtained by testing with multiple feature vectors, is used to compute a score χ defined as:

$$\chi = \eta(P(k_i \mid X) - \beta) \tag{4}$$

where η and β are constants for tuning the scale and the offset of the produced score.

A speaker-independent threshold θ, computed to minimize the decision cost, is then applied to the score (4) and a final decision $O(\theta)$ is made:

$$O(\theta) = \begin{cases} 1 & \text{for } \chi \geq \theta \\ 0 & \text{for } \chi < \theta \end{cases} \tag{5}$$

When the score χ is above or equal to the threshold the user claim is accepted, otherwise the utterance is considered to belong to an impostor speaker.

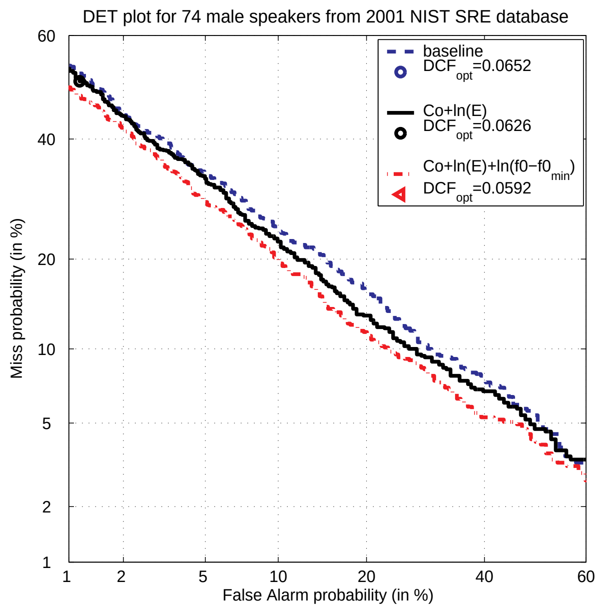

FIGURE 2. Speaker verification performance comparison for the baseline and the two modified feature sets

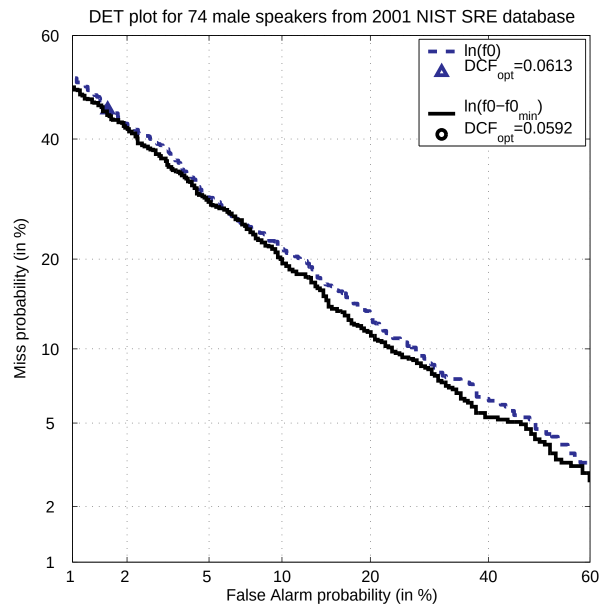

FIGURE 3. Speaker verification performance comparison for two representations of the fundamental frequency f0

2001 NIST SRE -- ONE-SPEAKER DETECTION DATABASE

The 2001 NIST SRE – one-speaker detection database was used in our experiments. It is excerpt from the Switchboard-Cellular corpora, which had been processed in order to remove any pauses and transmission channel echoes. The training data consists of spontaneous speech from 74 male and 100 female speakers, recorded in different environmental conditions: {‘inside’, ‘outside’, ‘vehicle’}. All training speech had been acquired over the mobile cellular networks of USA. Each target user is represented by about 2 minutes of spontaneous speech, extracted from a single conversation. The test data consist of speech recorded over {‘TDMA’, ‘CDMA’, ‘Cellular’, ‘GSM’, and ‘Land’} transmission channels. Both same and different phone number calls (implying different handsets) and different transmission channels are available for each user. Depending on the amount of speech the test trials contain, they are separated in the following five categories: {’00-15’, ‘16-25’, ’26-35’, ’36-45’, and ’46-60’} seconds. The complete one-speaker detection task includes all test trials, and therefore covers all aforementioned sources of variability. A comprehensive description of the evaluation database, and the speaker recognition evaluation rules, is available in the 2001 NIST SRE Plan [4].

EXPERIMENTAL RESULTS

All experimental results presented in this section, are only for the male part of the 2001 NIST SRE database. Our system has proved not to make significant differentiation between male and female speakers [5], and therefore we omit the female DET plots for simplicity of our exposition. The conclusions we derived from the male experiments, are valid for the female ones, too.

In all experiments, the training and the testing data sets have been processed in the way described in section “Preprocessing and feature extraction”. Approximately 40 seconds of voiced speech were available for training the male PNNs. Then, these PNNs were examined by testing with all speech trials (belonging to the specific gender), as defined in the complete one-speaker detection task. The male control file ‘detect1.ndx’ includes 850 target and 8500 impostor trials with length from 0 to 60 seconds of speech, and all available transmission channels.

The DET plots shown in Fig.2, Fig.3, and Fig.4, represent results obtained for the male users in the complete task. The marks ‘circle’ and ‘triangle’, point to the optimal value of the decision cost function designated as DCFopt. For some of the experiments, the optimal decision point lies outside the visible area, and therefore the mark symbols cannot be seen. The optimal values of the DCFopt are shown in the corresponding figure legend table.

Fig.2 presents comparison of the SV performance for the baseline and the modified feature sets, when user codebooks composed of 128 vectors and an UBgCB composed of 256 vectors are considered (which corresponds to their

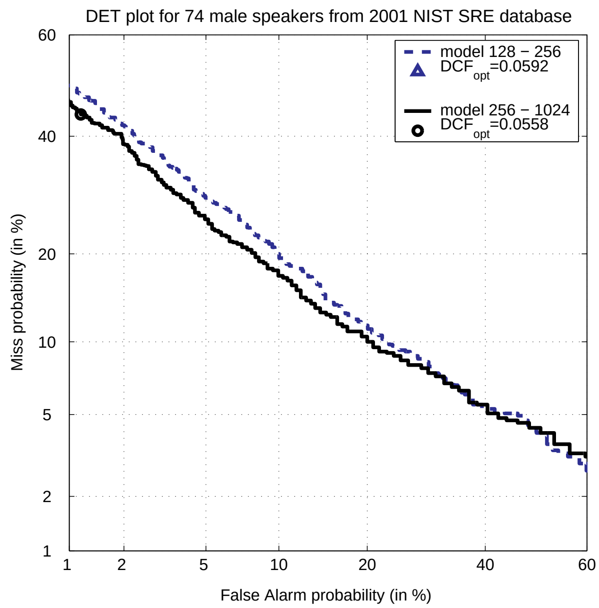

FIGURE 4. Speaker verification performance in dependence on the user's and the background codebooks size

size as it is defined in the baseline WCL-1 system). With dashed line, the SV performance obtained for the baseline MFCC features is depicted, while the solid line represents the case when the first MFCC is replaced by the logarithm of the frame energy, noted as ln(Energy). The dash-doted line reveals the performance for a feature set composed by adding a logarithm of the difference between the fundamental frequency f0 and a constant $f0_{min}$, to the latter feature set. As Fig.2 shows, a decrease of the error rate is observed when the first MFCC is replaced by the logarithm of the frame energy. It is well-known that the first MFCC is sensitive to the transmission channel and handset characteristics, and thus susceptible to unwanted variations. The logarithm of the frame energy is added to the rest of the MFCCs to compensate for the speaker-dependent information, lost with removing of the first MFCC. More significant reduction of the error rate is observed, when a logarithm of the fundamental frequency (minus a constant $f0_{min}$=55 Hz) is also added to the logarithm of the frame energy and the MFCCs – see the dash-doted line. The fundamental frequency carries important information about the specific glottal anatomy of the speaker, and together with the frame energy describes the prosodic style of the speaking person, which is important cue in the SV process. The dynamic range of the logarithm of the fundamental frequency however does not represent well its relative importance to the other parameters in the feature vector. Therefore, to solve this problem, we make use of the logarithm of the difference between the fundamental frequency and the constant $f0_{min}$. In Fig.3, a comparison between the cases when the speech feature vector contains ln(f0) (dashed line) and ln(f0-$f0_{min}$) (solid line) is shown. The noticeable reduction of the error rates and the DCFopt values for the case of ln(f0-$f0_{min}$) is obvious.

Fig.4 presents DET plots, obtained for different sizes of the user and the background codebooks, when the best feature set: {'ln(f0-$f0_{min}$)', 'ln(Energy)', 'MFCC(2:32)'} is considered. The baseline result, plotted with dashed line represents the case when 128 vectors are used for the user codebook, and 256 vectors for the background model. The solid line depicts the year's 2003 WCL-1 system, in its final configuration, with 256 vectors for the users' codebooks and with 1024 vectors for the reference background codebook. The reduction of the error rate and the optimal decision cost is clearly visible. When the last result is compared to the one obtained from the baseline 2002 WCL-1 system in its original configuration (only MFCCs, 128 vectors for the users' and 256 vectors for the background codebooks), a reduction of the equal error rate from about 18.5% to about 13.8% is observed, which corresponds to a relative reduction of the error by more than 25%.

CONCLUSIONS

An improved version of our baseline WCL-1 system, which is going to participate in the 2003 NIST SRE, was presented. Enhancements in the feature extraction and speaker modelling stages were performed. Improvements in the feature extraction step are connected to: reduced influence of the transmission channel and the handset variations over the presentation of the speech spectrum, and to adding prosodic features, which account for the speaking style of the users. At the modelling stage, refining the quality of the user and the background codebooks were preformed, by introducing pre-initialization of the k-means clustering algorithm with a small subset, randomly selected over the user training data, and more importantly by exploiting larger codebooks allowing more accurate representation of the users' and the reference models. The combined effect of all improvements comprises a reduction of the absolute error rate by about 4.7% at the equal error rate point, which corresponds to a relative reduction of the total error by more than 25%.

ACKNOWLEDGEMENTS

This work was supported by the "Generic Environment for Multilingual Interactive Natural Interfaces" - GEMINI project (IST-2001-32343).

REFERENCES

[1] Specht, D. F., “Probabilistic Neural Networks”, *Neural Networks*, Vol. 3, No.1, 1990, pp. 109-118

[2] Rabiner, L. R., Cheng, M. J., Rosenberg, A. E., McGonegal, C. A., “A Comparative Performance Study of Several Pitch Detection Algorithms”, *IEEE Transactions on ASSP*, Vol. ASSP-24, No.5, 1976, pp. 399-418

[3] Hartigan, J. A. and Wong, M. A., “A k-means clustering algorithm”, *Applied Statistics*, No.28, 1979, pp.100-108

[4] NIST, “The NIST Year 2001 Speaker Recognition Evaluation Plan”, March 1st, 2001, Available: http://www.nist.gov/speech/tests/spk/2001/doc/2001-spkrec-evalplan-v05.9.pdf

[5] Ganchev T., Fakotakis N., Kokkinakis G., ”Text-Independent Speaker Verification: The WCL-1 System”, *An International Conference on Text Speech and Dialogue TSD 2003*, Ceske Budejovice, Czech Republic, September 8 - 11, 2003. (accepted paper)