FRAME SELECTION FOR SPEAKER IDENTIFICATION

Maïa E.M. Weddin

GN ReSound A/S
Taastrup, Denmark
maiwed@gnresound.dk

Ass. Prof. Ole Winther

Technical University of Denmark
Lyngby, Denmark
owi@imm.dtu.dk

ABSTRACT

A novel approach to automatic speaker identification is presented. Using low-level acoustic feature sets, a frame based analysis of the system performance is implemented to locate the areas of a speech signal that contain a high level of speaker dependent information. Subgroups of frames are used for the text independent speaker identification task and the resulting system performance is compared with that of using all available frames. It is found that by exclusively using the frames in the transient areas of a speech signal, where the signal shifts between being voiced and unvoiced, the rate of correctly classified frames is increased by up to 14% compared to the case of using randomly selected frames. These results are obtained for PLPCC feature sets extracted from clean speech.

1. INTRODUCTION

Text independent speaker identification has been the focus of growing research interests over the past few years and new methods that aspire to decrease the error rate of these systems are constantly being developed and tested. The automatic speaker identification task is divided into three steps:

- Preprocessing
- Feature Extraction
- Speaker Modeling/ Classification (train/test).

The performance of the overall system is dependent on each of the three factors listed above, independently and combined. Selecting a feature extraction method and classifier often depends on the available resources and the intended application of the speaker identification system. There is to date no universally optimal feature set for use in speaker identification, and so the search continues to determine and extract features that contain a high level of speaker dependent information, thus enabling a classifier to more easily distinguish between different speakers.

Another property of the ideal feature set is that it is reliably obtainable and computationally feasible, even in situations where data is sparse. This is the driving force behind the widespread popularity of the short term spectral features that model the characteristics of the vocal tract, such as the Linear Prediction (LP) coefficients. These features are obtained through straightforward calculations that result in approximations of the speech envelope. These can then be transformed into cepstral coefficients which are commonly applied in state of the art systems [1].

The adequacy of a feature set for speaker identification lies in its ability to model signal properties that are unique for each speaker. Human speech perception is capable of quick and unambiguous classification of human voices and so feature extraction methods can be enhanced by approximating the auditory processes that take place physiologically, hence the advent of such feature sets as the Mel-frequency cepstral coefficients, MFCC, and perceptual linear prediction cepstral coefficients, PLPCC [2]. The MFCC feature set combined with a Gaussian mixture model (GMM) classifier has been shown to be highly effective, thus proving that the cepstral coefficients are not only computationally feasible, but that they are also significantly well suited for the speaker identification task [3]. These features, however, only model the vocal tract characteristics that filter the source signal formed at the glottis. As the latter can be assumed to be uncorrelated with the former, several source based feature sets have been created, motivated by the possibility of introducing complementary information to the feature space. Generally, these source features used in isolation yield less satisfactory results than their short-term spectral counterparts, however performance of a system can be improved by combining the two types of features [4],[5],[6].

The added demands on the amount of training and testing data required to extract high-level feature sets, as well as the increased dimensionality of the classifier when feature sets are combined, makes it desirable to determine an alternative means by which to increase system performance. In order to do this we will focus on the low-level features. These are usually extracted from 20-30ms frames, within which the speech signal is assumed to be stationary. Applying all frames to the speaker identification task is not necessarily the optimal implementation of these features, as the frames from different regions of the speech signal contain varying levels

of speaker dependent information. By heavily weighing only those frames that are rich in such content, redundancy and ambiguity within the feature set could be decreased and the rate of identification thereby increased.

It is therefore proposed that the speaker identification task could benefit from an additional step - one of *frame selection*.

2. FRAME SELECTION

The purpose of introducing frame selection is to prioritize frames from those areas of the speech signal that contain high levels of speaker dependent information. In the classification step, the identification decision is often based on the probability that a frame belongs to a certain speaker, so that speaker i is identified if $p(i) > p(j), i \neq j$. An entire sequence of test frames, X, can be classified by using the product of probabilities for N individual frames, as is done in GMM classifiers, see Eq.(1). The speaker i^* that maximizes this product is then selected as the owner of the voice.

$$p(X|\lambda_i) = \prod_{n=1}^{N} p(x^n|\lambda_i) \tag{1}$$

Alternatively, consensus can be applied. Using some form of decision logic, a classifier identifies each frame as belonging to a certain speaker and when this process is completed, the speaker that scores the largest share of the classified frames is identified as the correct speaker, so that instead of using Eq.(1) the ratio $\frac{n_i}{N}$, where n_i is the number of frames classified as belonging to speaker i and N is the total number of frames, is used.

Increasing the number of frames that are correctly assigned would increase reliability in speaker identification systems and eventually also allow for shorter test sequences to be classified correctly. In order to quantify the amount of frames that are correctly classified when a test sequence is analyzed, the performance metric that will be used here is not the traditionally implemented Equal Error Rate (EER), but rather the percentage of correctly classified frames, $\frac{n_{i,correct}}{N} * 100$.

The vital first step in frame selection is defining the criterion for the sorting of frames. This is initialized by labelling each frame in a sentence as being voiced or unvoiced. In applications involving features such as pitch estimates, only the voiced frames are needed [7]. Cepstral coefficients are extracted from both voiced and unvoiced frames and for the LPCC, MFCC and PLPCC feature sets, performance showed no signs of improvement when using either the voiced, or unvoiced, frames independently [8].

Voicing information is thus not sufficient to model a speaker's unique speech-producing vocal tract characteristics. Promising developments in the field of speech recognition have been obtained by pinpointing changes in the phonetic energies of the signal, such as the onset of a syllable [9]. Extending these findings to a less specific case in the hope that this can be applicable to speaker identification, transient areas of energy, indicated by the transition of a speech signal between a voiced and an unvoiced state, will be analyzed. In order to establish whether frames from these areas contain an increased level of speaker dependent information, they are selected from the full feature sets and tested independently. Due to the limited duration of these frames, several frames just after or immediately prior to a transition are also included in the analysis.

3. FEATURE SETS

The frame selection process does not replace any of the steps in the speaker identification system. It comprises of a modification so that the process includes four steps:

- Preprocessing
- Feature Extraction
- Frame Weighing
- Speaker Modeling/ Classification (train/test)

The frame selection and subsequent weighing is a process implemented independently from the feature extraction process and is thus not correlated with the feature set chosen. The nature of the frame selection criterion, however, requires that the features chosen be capable of capturing the change in signal energy that occurs in the transient regions. This is a dynamic property and thus such feature sets as the Δ and $\Delta\Delta$ derivatives of the cepstral coefficients are probably appropriate. These register the temporal change within a signal by averaging over the coefficients determined for several frames, as shown in Eq.(2), where $c_m(n)$ is the m^{th} cepstral coefficient for the n^{th} time frame. Θ is the number of frames that are included in the calculation.

$$\Delta c_m(n) = \frac{1}{\Theta}(c_m(n+\Theta) - c_m(n-\Theta)) \tag{2}$$

The $\Delta\Delta$ coefficients are then derived by applying Eq.(2) for the Δ coefficients.

4. THE SPEECH CORPUS

The trials here are implemented to solve the text independent speaker identification task for a closed set of 6 speakers, 3 men and 3 women. The speakers are taken from the ELSDSR database that was created by Ling Feng at the Technical University of Denmark in 2004 [10]. The speech is recorded in a quiet environment and with the same recording setup in each case. Each speaker contributes with the same 7 sentences for the training data set, and with two different testing sequences.

5. EXPERIMENTAL RESULTS

Preliminary trials without implementing frame selection and using MFCC, LPCC and PLPCC feature sets show that the highest level of correctly classified frames is repeatedly obtained for the PLPCC feature sets [8]. This can be contributed to the fact that these features approximate the auditory frequency analysis that takes place in the human ear prior to the linear prediction analysis, thus modeling speech as it is perceived physiologically and placing emphasis on those frequency regions that the ear is naturally more sensitive to. The experiments including frame selection are therefore implemented using PLPCC feature sets.

A nonlinear neural network is ideal for modeling complex data representations, and so a perceptron with a single nonlinear hidden layer consisting of 15 units is used as the classifier. The number of input neurons correponds to the dimensionality of the feature set and there are as many output neurons as there are reference speakers, i.e. 6 for these experiments. The output of each output neuron is transformed into a probability, y_j for the j^{th} neuron, subsequently classifying the speaker as being the one associated with the largest probability. Each test frame is paired with a target frame so that a calculation of the total percentage of correctly classified frames is possible. For each experiment, performance Z is measured as the rate in percentage of all correctly classified frames, i.e. over all $I = 6$ speakers in the set, as shown in Eq.(3):

$$Z = \frac{100}{N} \sum_{i=1}^{I} n_{i,correct} \tag{3}$$

Each training and test sentence is divided into short term frames that are 30ms in length and with 10ms overlap, and windowed so as to prevent distortions at frame boundaries. The preprocessing stage also includes preemphasis of higher frequencies with a first order high pass filter. The PLPCC coefficients of orders 9,11 and 13 and their Δ coefficients are extracted for each frame, as also determined.

The frame selection step then divides the frames into subgroups. First, each frame is labelled as being either voiced or unvoiced. This is done by calculating the autocorrelation function for each frame and then determining whether the signal block is periodic (or pseudo periodic), or nonperiodic. The nonperiodic frames signify the lack of fundamental frequency information and these frames are labelled as being unvoiced. The PLPCC feature sets are then split into the following subgroups:

1. UV1: This set is comprised of all the voiced frames that occur after an unvoiced frame, and all the unvoiced frames that directly precede a voiced frame.

2. UV2: This set includes all the voiced frames that are located just after a voiced frame found after an unvoiced frame, and the voiced frames preceding the voiced frames that occur directly before an unvoiced frame.

3. UV3: Just as in UV2, only the 3^{rd} voiced frame after or before an unvoiced frame, when the frames separating them are voiced.

Feature Set	*Random*	*UV1*	*UV2*	*UV3*
9PLPCC	51%	56%	57%	58%
9ΔPLPCC	45%	58%	61%	57%
11PLPCC	53%	58%	60%	60%
11ΔPLPCC	46%	58%	53%	57%
13PLPCC	52%	60%	63%	60%
13ΔPLPCC	49%	63%	62%	58%
13$\Delta\Delta$PLPCC	55%	67%	68%	68%

Table 1. Results for the PLPCC feature sets

Each subgroup includes frames from both the transitions from voiced to unvoiced and from unvoiced to voiced states, as the direction of the transition yields no observable difference in performance [8].

The neural network is trained and tested with 6s and 4.5s of speech from each speaker, respectively. For training, frames are randomly chosen either from the entire signal (*Random*), or from one of the listed subgroups, *UV1*,*UV2*, or *UV3*, so that the order in which the frames occur is changed for each experiment. The amount of training data is limited to 6s per speaker as it is constrained to the size of the smallest of the subgroups of features. No weighing scheme is implemented as the purpose of these trials is simply to determine whether isolating the frames that contain a transition between voiced and unvoiced speech, and those frames bordering such a transition, leads to an increase in correctly classified frames when compared with the random case, where frames are selected from anywhere along the complete signal.

The results are shown for the 9^{th}, 11^{th} and 13^{th} order PLPCC feature sets and their first order temporal derivatives in Table 1.

From Table 1, it can be seen that the highest rates of correctly classified frames are recorded for the 13^{th} order PLPCCs. Feature sets of higher order are capable of modeling finer details of the formant frequencies and it is interesting to note that this aids the identification rate for almost all of the frame subgroups. The results also indicate that the amount of correctly classified frames does increase when frames from the transient regions are selected and only the corresponding feature coefficients used. There is very little difference between using *UV1*, *UV2* and *UV3*, which signifies that the transition between voiced and unvoiced frames lasts for the space of at least 3 short-term frames.

No improvement in performance is registered for the Δ feature sets, so the $\Delta\Delta$ coefficients are derived for the 13PLPCC feature set, in the hopes of further increasing the correct frame

rate. This does indeed result in higher frame classification rates for all frame subgroups, the highest being for the *UV* sets, as before. The greatest measured improvement when using the transient frame groups is obtained for the Δ and $\Delta\Delta$, coefficient sets. This indicates that these feature sets are more adequate at modeling the information contained within the shifting areas of speech than the stationary coefficients are. The maximum benefit of using the *UV* subgroups is obtained for the 9ΔPLPCC and the 13ΔPLPCC feature sets, where an increment of 14% is observed.

The process of frame selection would be considerably more efficient if the transient areas could be located by analyzing the feature set without needing to first determine the voiced and unvoiced state of each frame. For the 13$\Delta\Delta$PLPCC feature set, it was observed that the summed absolute magnitude of the coefficients for the frames in the subgroups accounted for 46% of the summed absolute magnitude for all frames, while the subgroups themselves only include 16% of the total number of frames. The implications of this, and the implementation of other analysis methods, have yet to be explored.

6. CONCLUSIONS AND FUTURE WORK

For a small set of speakers and using speech not contaminated by noise, it has been shown that the number of frames used in a feature set for speaker identification can be reduced and yet the level of speaker dependent information increased by dividing each signal into subsets based on the transient areas of speech and then only using these subgroups of frames. Using up to 3 frames either after the transition from an unvoiced frame to a voiced one, or 3 frames prior to the transition from a voiced frame to an unvoiced frame, leads to increased performance of up to 14% in correct frame rate.

Future directions include generalizing these findings for larger speech databases and for longer training and testing sequences. Also, robustness in the case of speech contaminated by noise must be tested. Trials should be implemented in order to determine just how many frames bordering a transition can be used. Finally, a way of automizing the process of selecting the transient frames without necessarily having to label frames as being voiced and unvoiced first needs to be derived to enable a weighing scheme that is applicable in a speaker identification system and thereby realizing the potential benefits of frame selection.

7. REFERENCES

[1] J.R. Deller, J.H.L. Hansen, and J.G. Proakis, *Discrete - Time Processing of Speech Signals*, IEEE Press, 2000.

[2] H. Hermansky, "Perceptive linear prediction (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[3] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[4] B. Wildermoth, "Use of voicing and pitch information for speaker identification," M.S. thesis, School of Microelectronic Engineering, Griffith Univeristy, Australia, 2001.

[5] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *ICASSP '01*. IEEE, 2001, vol. 1, pp. 364–367.

[6] D. Reynolds, W. Andrews, J. Campbell, J. Navrátil, B. Peskin, A. Adami, Q. Jin, D. Kluá cek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "Supersid workshop: Exploiting high-level information for high-performance speaker recognition," in *ICASSP '03*. IEEE, 2003, vol. 4, pp. 784–787.

[7] A.G. Adami, R. Mihaescu, D.A. Reynolds, and J.J. Godfrey, "Modeling prosodic dynamics for speaker identification," in *ICASSP '03*. IEEE, 2003, vol. 4, pp. 788–791.

[8] M. Weddin, "Speaker identification for hearing intruments," M.S. thesis, Department of Informatics and Mathematical Modelling, Technical University of Denmark, 2005.

[9] S. Wu, M.L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *ICASSP '97*. IEEE, 1997, vol. 2, pp. 987–990.

[10] L. Feng, "Speaker recognition," M.S. thesis, Department of Informatics and Mathematical Modelling, Technical University of Denmark, 2004.