NEW CEPSTRUM FREQUENCY SCALE FOR NEURAL NETWORK SPEAKER VERIFICATION

Paul Cristea and Zica Vâlsan

"Politehnica" University of Bucharest
Splaiul Independentei 313, 77206 Bucharest, Romania
pcristea@dsp.pub.ro, zica@helix.elia.pub.ro

ABSTRACT

The influence of cepstrum parameters on text-dependent speaker verification and speech recognition is investigated. Experiments are performed to establish the relevance of various resonant frequencies and frequency bands in terms of their speech and speaker recognition ability. A Romanian database of eighteen isolated words has been used. The study of the filter bank analysis suggests a new frequency scale instead of the currently used mel-scale to extract from the speech signal cepstrum coefficients. The proposed scale results in better performance in speaker verification. The processes of speech recognition and speaker verification are carried out by using a neural network system comprising a self-organizing feature map (SOFM) and a multilayer perceptron (MLP).

1. INTRODUCTION

Speaker recognition is the process of automatically recognizing the person who is speaking on the basis of individuality information in speech waves. This technique is used to verify the identity of a person accessing a system. These services include but are not limited to banking transactions over a telephone network, telephone shopping, database access services, information services, voice-mail, security control for confidential information areas, and remote access to computer. There are two types of speaker recognition systems: speaker identification and speaker verification. Speaker identification determines from which of the registered speakers a given utterance comes. Speaker verification is the process of accepting or rejecting the claimed identity of a speaker. The fundamental difference between identification and verification modes is the number of decision alternatives. In the identification mode the number of decision alternatives is equal to the size of the population, whereas in the verification mode there are only two alternatives, accept or reject the identity claim, regardless of the size of population. Most applications of speaker recognition are actually speaker verifications. In order to avoid the possible breaking of the system by playing back a recorder voice of the target speaker the text-prompted method in which the speaker is requested to utter a key sentence show on display, is used. The system verifies the speaker's identity only when it decides the claimed speaker has pronounced the requested sentence.

In this study, the influence of cepstrum parameters on the result of text-dependent method speaker verification and speech recognition is investigated. We present a study on the frequency analysis for the problem of speaker verification and speech recognition. Consequently, the validity of using the mel-scale for parameterization in speaker verification is questioned and a more appropriate scale for speaker verification is proposed. In order to compare the performance of the new frequency scale with that of the mel-scale, an experiment was performed on a Romanian isolated words database. The results are presented in the last section.

2. FEATURE SET

The human perception of the sound frequency content, either for pure tones or for speech signals, does not follow a linear scale. Research on this matter has lead to the idea of defining subjective pitch or pure tones. Thus, for each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called the "mel scale". As a reference point, the pitch of a 1KHz tone, 40dB above the perceptual threshold, is defined as 1000 mel.

The mapping between the perceived frequency scale (mel) and the real frequency scale (Hz) is given by the relation

$$F_{mel} = 2595\lg\left(1+\frac{f_{Hz}}{700}\right) \qquad (1)$$

Loosely speaking, it has been found that the perception of a particular frequency by the auditory system is influenced by the energy in a critical band around that frequency. In this idea the mel scale-spectrum is simulated using a filter bank spaced uniformly on a mel scale, where the output energy from each filter band approximates the modified spectrum. Fig.1. shows an example of such a filter bank, in which each filter has a triangular band-pass frequency response, and the spacing as well as the bandwidth are determined by a constant mel frequency interval. [1]. Speech features derived using the mel

0-7803-5682-9/99/$10.00©1999 IEEE.

scale have also resulted in superior speech recognition performance when compared to parameters obtained with a linear frequency scale [2].

In this study, our main point is that the problem of speech recognition is different from the problem of speaker verification in what concerns the frequency scale. The speaker characteristic vector has to contain information related to the source of the vocal signal and the structure of the vocal tract. Pitch and delta pitch are used as typical parameters of the source, while cepstrum and delta cepstrum are used as typical parameters for the vocal tract. Although the difference in pitch between speakers is large, it is difficult to effectively use the pitch for speaker recognition, as people can easily change the pitch of their voice. Cepstrum and delta cepstrum are better for his purpose as being rather invariant for a given speaker. Usually, a non-native speaker tries to imitate a native speaker as closely as possible, attempting to correct perceptually the most significant differences in pronunciation with respect to the native speaker pronunciation. Therefore a parameter set which is based on perceptual criteria is not adequate for the problem of speaker verification.

3. EXPERIMENTS USING FORMANT FREQUENCIES

We have performed some experiments in order to assess the relative significance of the formant frequencies for both above-mentioned problems. The database contains eighteen Romanian isolated words (20 speakers x 18 words x 15 tokens). Voiced sections of each word in the database were extracted like in [3]. For each vowel, the first three formant frequencies F1, F2, F3 were detected and their derivative were estimated using a 20 ms Hamming window shifted every 10ms.

A neural network, that comprises a SOFM (2 inputs, 4 x 4 feature map) and of a MLP (16 x 20 x 2 neurons), was used for each formant F*i* and its derivative delta F*i* from the database (i.e., for F1 and delta F1). The system evaluates the speaker verification performance using the formant structure. The node in the map that has the weight vector closest to the input vector, i.e., the winner, gives the response and is assigned the value one, while all the other nodes are set to the value zero. A matrix of the same dimension as the feature map is used to accumulate these outputs. After the end of an input word, a vector made by cascading the columns will be the input in the MLP. The MLP is trained with the backpropagation algorithm using an adaptive learning rate [4]. The results about the influence of formant frequencies on the performance of speaker verification are shown in table 1.

In a second experiment we have used the same type of neural system comprising a SOFM (4 x 4 neurons) and a MLP (16 input x 20 hidden x 18 output neurons) and the same input parameters as those mentioned above, but we focussed on evaluating the speech recognition performances. These performances for each formant are shown in table 1. An interesting question is whether all phonemes are useful for speaker modeling, or whether is better to ignore some of them when doing speaker verification. The experiment has shown that all vowels - including the "ill-reputed" back vowels (/o/ and /u/) - are useful for speaker verification and contribute in reducing the global error rates.

When speaker verification and speech recognition performance are compared, F2 was found to be the most significant resonant frequency contributing to correct classification for both speech and speaker recognition. As already known, the F1 formant resulted to be the most important for the speech recognition problem, but almost irrelevant for speaker verification.

Formant frequencies	Speaker verification rate	Speech recognition rate
F1	30.2%	53.1%
F2	43.5%	70.2%
F3	32.1%	41.3%

Table 1. The influence of formant frequencies on the performance of speaker verification and speech recognition.

4. EXPERIMENTS USING FILTER BANKS

In order to investigate the speaker verification and speech recognition ability of various frequency bands, we have performed two experiments starting from the aspect of sound perception using the same neural network system and database as in the previous experiments. This time, instead of the formants, the cepstrum coefficients were used as input.

For speaker verification the experiment has been conducted as follows. A local database, which contains the tokens for each word, over the acquisition period, is created for each speaker. The reunion of these individual databases for all speakers generates the initial global database of the system. The audio signal was sampled at 22KHz with an 8bit digitizer. Test and training data are analyzed at 20 ms frame interval with a Hamming window shifted every 10ms, using a simulated critical band filter-bank. The 0.1 - 4 kHz frequency range was divided into 20 bands equally distributed on the mel scale. The 20 mel frequency cepstrum coefficients characterizing a frame of speech are calculated using the critical band filters in fig.1, where:

F$_c$[Hz]	174	250	335	425	524	635	754	942	1052	1190
SVR%	67.2	68.3	69.5	68.6	70.6	70.1	71.3	72.5	73.1	73.6
SRR%	60.3	79.6	80.1	87.3	90.1	93.2	94.5	95.2	96.3	97.8

F$_c$[Hz]	1347	1523	1718	1930	2161	2414	2688	2986	3311	3664
SVR%	74.2	88.3	90.5	96.2	97.1	94.1	84.3	70.2	76.1	64.1
SRR%	99.4	99.1	96.6	85.3	80.1	78.2	70.2	60.2	52.1	41.6

Table 2. Comparison of speaker verification versus speech recognition performance.
F$_c$ - Center Frequency of filter bank, SVR - Speech Verification Rate, SRR - Speech Recognition Rate.

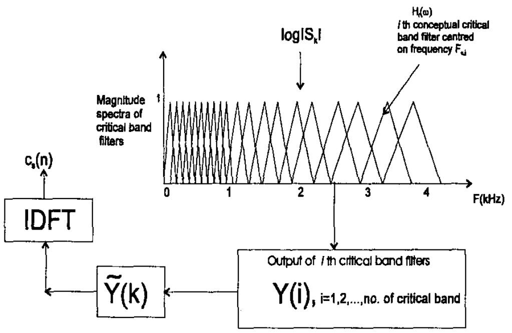
H$_i$(ω)
I th conceptual critical band filter centred on frequency F$_{ci}$
log|S$_k$|
Magnitude spectra of critical band filters
1
0
1
2
3
4
F(kHz)
c$_s$(n)
IDFT
Output of *I* th critical band filters
Y(i), i=1,2,...,no. of critical band
$\tilde{Y}$(k)

Figure 1. Use of critical band filters to compute the mel-cepstrum.

$Y(i)$ is the weighted log energy in the ith critical band,

$$Y(i) = \sum_{\substack{\text{Small range of} \\ k\text{'s around } k_i \\ k_i \propto Fc_i}} \log|S(k)| H_i\left(k\frac{2\pi}{N'}\right) \tag{2}$$

$$\tilde{Y}(k) = \begin{cases} Y(i), & k = k_i \\ 0, & other\, k \in [0, N'-1] \end{cases} \tag{3}$$

N' is the number of points used to compute the short time Discrete Fourier Transform,

$c_s(n)$ are the mel frequency cepstrum coefficients,

$$c(n) = \frac{2}{N'} \sum_{i=1,2,,\ldots,N_{cb}} \tilde{Y}(k) \cos\left(k_i \frac{2\pi}{N'} n\right) \tag{4}$$

n is index of mel frequency cepstrum coefficients, running from 1 to 20 in our experiment,

N_{cb} is number of critical bands.

The mel frequency cepstrum coefficients are computed and presented at the input of the system. The neural network has the following structure: SOFM with 20 input neurons (equal to the size of mel frequency cepstral coefficient vector) and (7 x 7) output neurons and a MLP with (49 x 50 x 2) neurons. After the training of the system for a word and a specific user, the model of that word spoken by the user is contained in the weights of the neural network. The operation is repeated for all the words uttered by the speaker, and the reference model for that speaker is created and stored. Again, the process is repeated for all registered speakers, so that the weights for each of them are stored in the global model. The classification performance using test vectors are shown in table 2. The rate of correct speaker verification is the best in the 1500-2500Hz-frequency range that corresponds to the F2-F3 range.

For the speech recognition experiment, the same pre-processing and neural system as described above have been used, except that the number of output neurons of MLP has been chosen equal to the number of words (18). The system is trained until all words from the global and local databases are correctly classified. The speech recognition performance in function of the frequency is also shown in table 2.

For speech recognition, the performance is better for lower frequencies than for mid-range frequencies. In conclusion, at least for Romanian speakers, the mid-range (1500-2500Hz) contributes more to speaker verification performance, whereas low frequencies are more important for speech recognition performance.

This result suggests a new frequency scale to achieve a better performance in the speaker verification case. A larger number of filter banks have been concentrated in the mid-range frequency domain, to emphasize the contribution of the better performing filters. The 20 center frequencies of the filter bank that range between 0-4 kHz are given in table3 and figure 2.

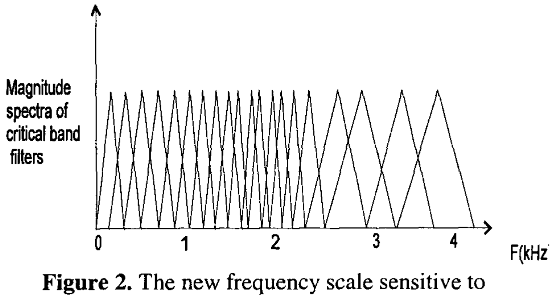

Figure 2. The new frequency scale sensitive to speaker verification

F_c[Hz]	250	390	682	794	958	1150	1300	1450	1600
F_c[Hz]	1750	1900	2150	2300	2414	2688	2986	3310	3664

Table 3. Center frequencies for the new frequency scale

5.SPEAKER IDENTIFICATION/REJECTION METHOD

Crucial to the successful operation of speaker verification systems is the maintenance of good reference models. This is important for speaker verification systems used for access control, where it can be expected that the user verification will take place at large time intervals of time, during which trial-to-trial variations are likely to occur. Reference models are constructed for each user in one or more training sessions, by selecting the models corresponding to the best identification rate. First, we collect and analyze relatively few samples of a new speaker's identification utterances in a controlled enrollment session. Subsequently, the user utterances from successful identification trials are used to update the speaker's reference models. Adaptation of reference models carried out in this way can provide comprehensive and robust models, which can accommodate reasonable trial-to-trial variations.

An issue related to this problem is the assignment of thresholds for identification trials. Thresholds must be assigned to tolerate trial-to-trial variations, at one hand, and to ensure a desired level of performance, at the other. A "tight" threshold makes it difficult for impostors to be wrongly accepted by the system, but at the risk of falsely rejecting legitimate users (customers) because of the long time variation of their characteristics. Conversely, a "loose" threshold enables customers to be accepted consistently, at the risk of accepting impostors

6. COMPARSION OF CEPSTRUM SCALES

In order to compare the performance of speech recognition with that of speaker recognition, using the new frequency scale, an experiment was performed on the database. Two sets of parameters were extracted:

First, ten mel frequency cepstrum coefficients and ten delta parameters (calculated as a polynomial expansion using a time sequence of 5 mel frequency coefficients over a period of 100 ms).

Second, ten cepstrum and delta cepstrum coefficients were established by using the new frequency scale.

Each set of parameters has been used as input to the neural network system described above. We have used the same parameterization approach for isolated words for both sets.

The average speaker verification rates for these parameters set is shown in Table 4. It can be observed that the new frequency scale performs better than the mel-scale for the speaker verification.

Comparison of different frequency scales in speaker verification		
	mel-frequency scale	new frequency scale
Speaker verification rate %	97.1	99.3

Table 4. Comparison of the mel scale and new frequency scale in terms of their speaker verification performance.

7. CONCLUSION

In this study we have investigated the influence of formant frequencies on the efficiency of speaker verification and speech recognition systems. A new scale has been proposed for establishing the cepstrum coefficients, based on the analysis of performance *vs* frequency for a Romanian isolated word database. The new scale resulted in a better performance for speaker verification than the mel scale.

8. REFERENCES

[1] W. Koenig, "A new frequency scale for acoustic measurements", *Bell Telephone Laboratory Record*, 27 pp. 299-301, 1949

[2] S. B. Davis, P. Mermelstein, "Comparison parametric representation for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust., Speech, Sig. Proc.*, 28(4), pp 357-366, Aug. 1980

[3] I. Gavat, B. Sabac, Z. Vâlsan, " Phoneme based text-prompted speaker identification with multilayer perceptron". *Proceedings of the Third International Conference on Technical Informatics, CONTI'98*, pp. 226-233, Timisoara, 29-30 Oct. 1998.

[4] Z. Valsan, B. Sabac, I. Gavat, "Combining self-organizing map and multilayer perceptron in a neural system for fast key-word spotting", *International Workshop Speech and Computer SPECOM'98*, pp.303-306, St. Petersburg, Oct. 1998.

[5] R. Mammone, "Robust speaker recognition", *IEEE Signal Processing Magazine*, pp.58-71, Sept. 1996.