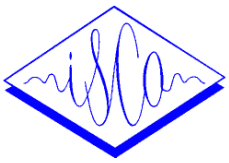

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

SPEAKER RECOGNITION EXPERIMENTS IN ESTONIAN USING MULTI-LAYER FEED-FORWARD NEURAL NETS

Toomas Altosaar & Einar Meister
Acoustics Laboratory
Helsinki University of Technology
Otakaari 5A, 02150 Espoo, Finland
Toomas.Altosaar@hut.fi

Laboratory of Phonetics and Speech Technology
Institute of Cybernetics, Estonian Academy of Sciences
Akadeemia tee 21, Tallinn EE0026, Estonia
einar@ioc.ee

ABSTRACT

In this paper a general strategy towards robust and efficient speaker recognition is presented. Emphasis is placed on comparing the usefulness of different features calculated from the speech signal at different temporal and spectral resolutions. Specifically, three spectral features are evaluated in a neural network environment: linear frequency loudness scaled spectra, auditory spectra from an auditory model, and the lattice coefficients from a warped linear predictor. These features are tested with four different neural network topologies ranging from speaker identification to verification configurations. Variations in the neural net dimensions are also performed to gain an understanding of the complexity of the problem. The tests are based on 40 minutes of speech recorded from a set of 20 native Estonian speakers.

1. INTRODUCTION

Although there exists a remarkable amount of knowledge and experience in Estonian phonetics, speaker recognition is a new research area being pursued in Estonia. A pilot project on text-independent speaker recognition was recently initiated and its objectives were to:

- find a set of reliable features that could be used for speaker recognition tasks
- implement feed-forward neural nets for classification of these features
- carry out preliminary experiments on a set of 20 speakers
- develop a hierarchical structure for a speaker recognition system

This paper presents the findings from a set of speaker recognition experiments where the usefulness of different features are measured. 279 different multi-layer feed-forward (MLF) neural networks were trained and their performance reported on with variations made to the temporal and spectral resolutions of different features, and with different network dimensions and topologies.

2. GENERAL STRATEGY & SYSTEM HIERARCHY

The general recognition strategy adopted in this paper has been that machine based speaker recognition should try to model the process carried out in human listeners. This means that a set of sequential decisions must be made, e.g., whether the sound is speech or noise, to which gender/age group the speaker belongs to, and then based on temporal and spectral cues begin eliminating certain candidates from the available set of speakers and matching towards a known speaker.

From this model it follows that the recognition process should be carried out hierarchically in several steps. This permits the analysis procedures that are applied initially to be more general and robust when compared to the features used and recognition methods utilized during later phases of the process. After each step fewer possible speaker candidates exist in the search space thus allowing more specific methods to be applied. Currently we are using only 3 different stages but due to the hierarchical structure it is possible to integrate new stages into the existing framework easily. We believe that such a hierarchical speaker recognition system should be effective as well as efficient. Figure 1 shows these ideas graphically.

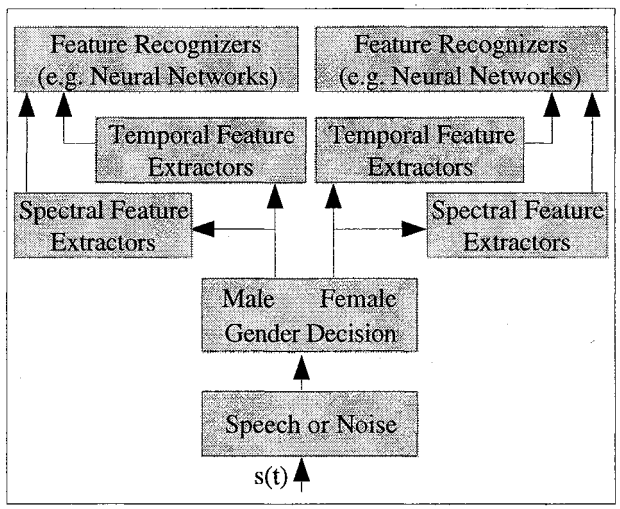

Fig. 1. General strategy and hierarchy of the proposed speaker recognition system.

3. SPEECH MATERIAL

A set of 43 sentences including all Estonian phonemes in different prosodic contexts was compiled. Approximately 4 minutes of speech was recorded in an office environment from each speaker (5 males and 5 females, speakers 1-10, odd = male, even = female) at a 12.5 kHz sampling frequency using a 16 bit A/D converter. In addition to this material one minute of speech from 10 additional speakers (5 males, 5 females) was also recorded. In total 540 utterances from 20 native Estonian speakers made up the database on which the experiments were based.

4. FEATURES

Effective speaker recognition requires the use of both temporal and spectral information derived from the speech signal in order to achieve high rates of performance.

4.1. Temporal Features

Stress and rhythm play important roles in speech production and are manifested as implicit low-level physiological activities learned early on in life. They can represent quite uniquely the characteristics of a speaker and for this reason we believe them to be important cues in the recognition process.

On the basis of average fundamental frequency (F0) speakers can be divided into the two gender groups. This immediately allows the pattern recognition modules, e.g., neural networks, to be more specialized in nature and may allow for improved performance levels to be achieved. F0 patterns, their distributions and shapes may also provide clues to the identity of the speaker. In these experiments F0 information was used to separate the speakers into male and female classes.

4.2. Spectral Features

Three different spectral measures were calculated from the signals. These were later used as inputs to the MLF nets and their discrimination potential measured. In all cases a 20 ms Hamming window (250 points) was first applied to the speech signal before the following types of spectra were calculated.

Level thresholding was also applied to avoid processing periods of silence or low-level ambient noise, as well as to reduce computational requirements. Therefore, spectra were calculated only at locations where the signal was in the top 80 % level of its loudness range calculated over an entire utterance. The simple loudness measure used was the following: a) a 20 ms Hamming window was applied to a speech segment, b) the windowed samples were then squared and accumulated together, c) the accumulated value was raised to the 0.25 power. This measure was calculated at 10 ms steps throughout each utterance.

Different frequency resolutions for the loudness and auditory spectral types were also calculated. This was done so as to gain a better understanding of how much data reduction could be performed before performance degradation occurred in the MLF nets.

4.2.1. Loudness Spectrum (LS)

The 256 point FFT power spectrum was loudness scaled, i.e., each element was raised to the 0.25 power. This compression was performed so that the MLF nets were provided with well scaled input values (normally between 0 and 2). The 128 frequency components were downsampled in the frequency domain to see whether the spectral resolution could be reduced without any loss of recognition performance. Downsample factors used were 2, 4, and 8 and produced 3 different LS features: a) a 64 point spectrum (denoted by LS64), b) a 32 point spectrum (LS32), and c) a 16 point spectrum (LS16).

4.2.2. Auditory Spectrum (AS)

The output of an auditory model [1] was chosen as another spectral feature. The Bark scale provides more frequency resolution at lower audible frequencies than at higher ones. The auditory spectrum was calculated at a 0.5 Bark resolution to yield 42 channels (denoted by AS42). Lower frequency resolution auditory spectra were also tested: 1 Bark (21 channels, denoted AS21), and 1.5 Bark (14 channels, denoted AS14).

4.2.3. Warped Linear Prediction (WLP)

A compact type of spectral measure was also chosen for the spectral feature comparison. The lattice coefficients from a 9th order warped (Bark frequency scale) linear predictive model were used [2]. The nine lattice coefficients were denoted as WLP and were tested to see how well they would perform against the larger LS and AS input features.

4.3. Long & Short Term Spectral Training

Both long and short term spectra were used as training material for the pattern recognition units. This was done to see how the MLF nets would react to highly averaged spectra (and consequently a low number of training elements) vs. short term spectra (where a much larger number of training elements existed).

4.3.1. Long Term Spectra

In order to examine the discrimination potential of the long term spectrum a preliminary test was performed with the Euclidean distance (ED) measure. The ED-s between the long term spectra of different utterances of the same speaker were frequently smaller than the ED-s between the long term spectra of the same utterances of different speakers. This confirmed that the long-term spectrum contains valuable speaker information and constitutes a simple but robust method.

Long term spectra were calculated only at significantly high speech energy levels, as mentioned above. Spectral slices were calculated at 10 ms intervals and averaged to yield a single spectrum for an entire utterance.

4.3.2. Short Term Spectra

Short term spectra were also experimented with. By supplying the MLF nets with a much larger number of training elements the generalization capability of the recognition units could be tested. These training elements were generated by calculating the averaged spectra at "islands of energy" locations within an utterance. An island constituted a contiguous area of the loudness curve where the level was in the top 80 % range. To avoid many very short term averages being generated an additional stipulation was added: the "island" had to be at least 50 ms in duration. This removal of short segments may also give the desired effect of eliminating parts of speech that do not contribute essentially relevant information to the recognition process (e.g., stops) while including more important indicators such as nasals and vowels [3].

4.4. Level Normalization

To avoid different relative signal levels in affecting the pattern recognition units, level normalization was performed on the spectral features. For the LS and AS cases the spectrum was scaled by a normalization factor dependent upon the loudness level within an utterance. For the WLP case level normalization was achieved by adding a 10th term to the lattice coefficient vector that indicated the relative level of the signal.

5. FEATURE RECOGNITION

The different features were presented to discriminative multi-layer feed-forward neural networks with a single hidden layer using a standard back propagation training algorithm.

5.1. Net Topologies

Four different net topologies were experimented with and are shown in figure 2.

- NT1 general identification net trained on all speakers
- NT2 identification net trained on all male speakers
- NT3 same as NT2 except for females only
- NT4 verification net trained to recognize only male No. 1

The nets were given one integrating output node for each speaker that was to be recognized [3]. An additional output node denoted by X was also created and a subset of the 10 additional speakers ("unknowns", numbers 1, 2, 3, 4, 5, and 10) was used to train it. By directing these "unknown" speakers to the X node the net was forced to identify these speakers as one. This was done so as to create more selective surfaces for the "known" set.

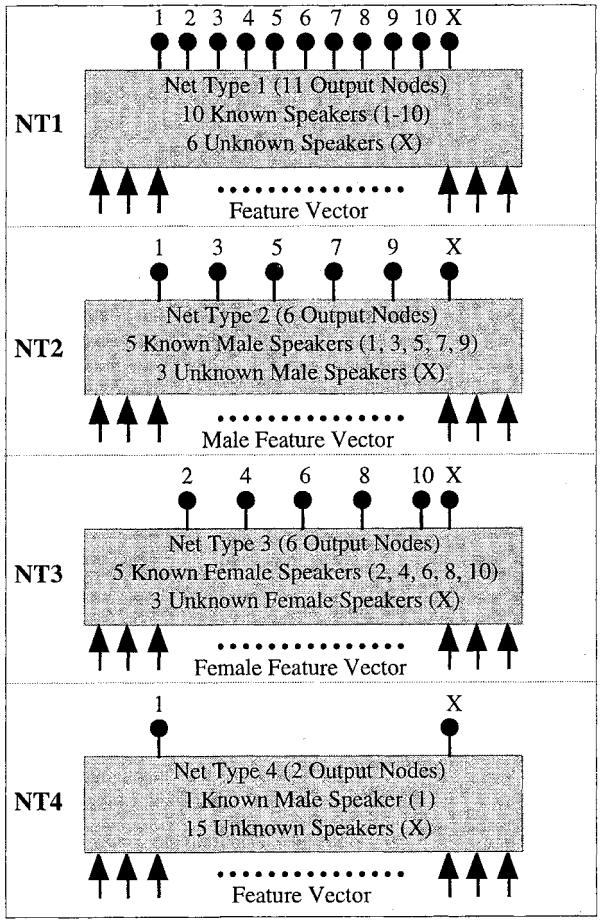

Fig. 2. Four different net topologies used in the speaker recognition experiments.

5.2. Training and Evaluation Sets

Spectral features were generated from the 540 utterances. From the 43 utterances of the larger speech set 29 sentences (≈2/3) were chosen to be included in the training set since they represented a phonetically diverse but representative data space. The remaining 14 sentences were used to generate the evaluation set. Thus the training material for the nets were identical sentences spoken by different speakers. To train the additional X node, 7 sentences out of 11 from the additional speakers were chosen (≈2/3). The remaining 4 sentences were included in the evaluation set.

For the long term spectra one training element was generated for each utterance as described above. Training element population normalization was carried out to better represent the "unknown" speakers. In total 464 training and 163 evaluation elements were used. The short term test had more training and evaluation elements by approximately a factor of one magnitude. Now the set sizes were ≈5000 for training and ≈2000 for evaluation. For net types NT2 and NT3 the training and evaluation set sizes were approximately half of the figures mentioned above due to gender selectivity.

5.3. Long Term Recognition

MLF networks were trained on the 7 different spectral representations: LS64, LS32, LS16, AS42, AS21, AS14; and WLP. The number of hidden nodes was also varied from 2 to 10, i.e., 9 different nets were trained on a single spectral representation. Evaluation was performed after every 25 presentations of the entire training set to the net. Training was terminated after 500 presentations of the training set. The state of the net when the performance was at a maximum was trended and is shown in the following figures. Each figure contains the results of 7x9=63 different nets.

Figure 3 shows the results of NT1 when trained on different long term spectra. Even with 6 hidden nodes most representations performed in the 85-90 % range. LS64, LS32 and AS42 performed well.

Figure 4 shows the results of NT2 when trained on different long term spectra from only male speakers. As is evident NT2 performed much better than NT1 since these nets were specialized for male speakers, and also the difficulty of the problem was smaller since now there were only 8 speakers involved. LS64, AS42, and AS21 performed well. Now 4 hidden nodes was sufficient for almost perfect recognition.

The results of the female NT3 nets are shown in figure 5. Performance levels are not as high as for NT2 — this may be due to the female voices being more similar to one another when compared to the similarity of male voices. LS32, AS21, and AS42 functioned well.

Results for recognizing male speaker no. 1 from a set of 15 other male and female speakers are shown in figure 6. Here the performance levels are the highest indicating that the net was able to specialize well. The best representations are: LS64, LS32, and LS16. Surprisingly, WLP performed well.

5.4. Short Term Recognition

A comparison was made with 27 nets trained on the larger short term auditory spectra sets at three different frequency resolutions. Since the number of training elements was ≈10 times larger, and that much more spectral variation was now present, the difficulty of the problem was expected to be much harder. Therefore, addition hidden nodes were allocated to the nets. Figure 7 shows the comparison: performance levels are much lower. Even with 27 hidden nodes the recognition rate does not exceed 85 %. Nearly perfect recognition is available with only 4 hidden nodes when using long term spectra.

6. SUMMARY

This paper compared the effectiveness of different spectral representations and resolutions among themselves given equivalent tasks. From these experiments it was determined that the loudness spectrum performed well even at half and

quarter resolutions. The auditory spectrum was also useful at full and half resolutions but at a 1.5 Bark resolution the performance was poor. The lattice coefficients from warped-LP fared poorly when compared to the LS and AS representations. This may be due to the nature of the representation: a small amount of noise can cause the reflection coefficients to take on much different values making their recognition much more difficult. Long term spectra was much more easily identified when compared to shorter term spectral averages. Higher performance can be achieved by specializing the networks. In these experiments it was seen that introducing gender nets improved performance substantially, and by specializing the net to one speaker even higher recognition rates were possible. The network's selectivity potential was forced to be high by assigning several speakers to a single output node.

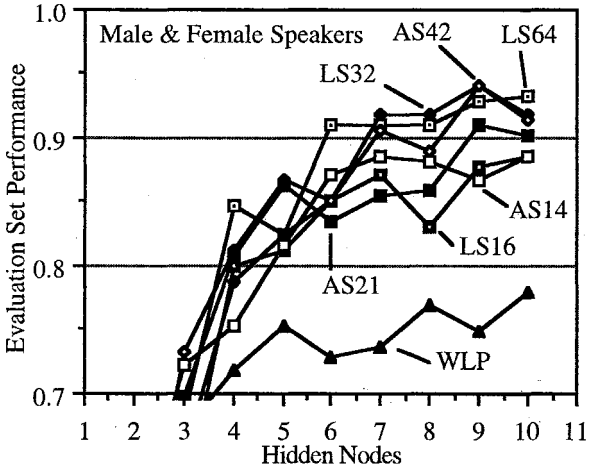

Fig. 3. Performance for general male and female speaker identification nets NT1.

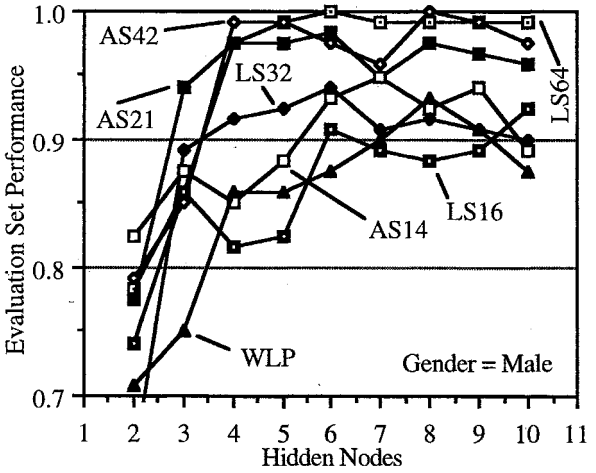

Fig. 4. Performance for specialized male speaker identification nets NT2.

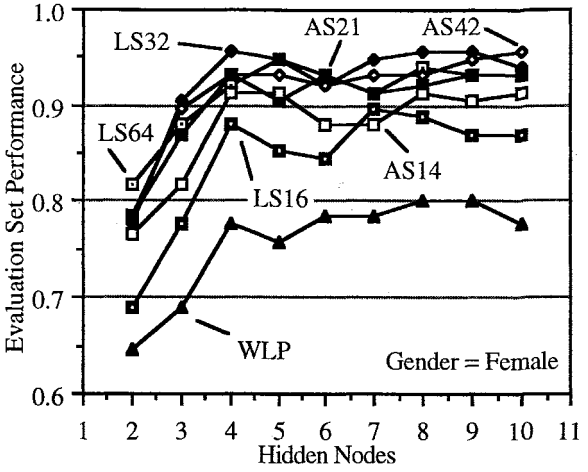

Fig. 5. Performance for specialized female speaker identification nets NT3.

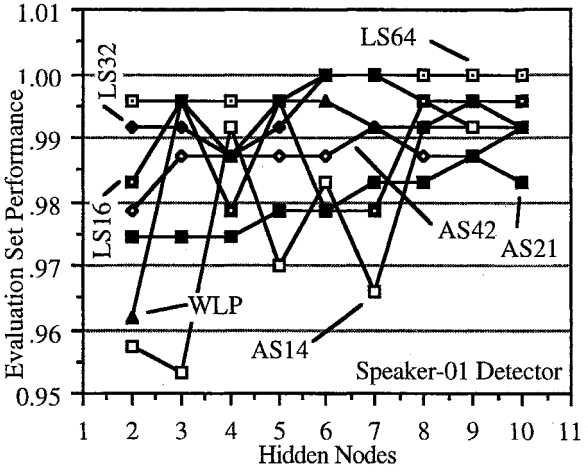

Fig. 6 Performance for specialized speaker no. 1 recognizer nets NT4.

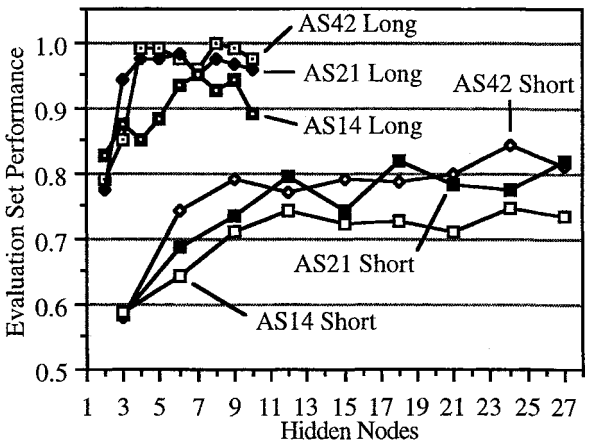

Fig. 7. Short vs. Long term spectral representation net performance for NT2.

7. REFERENCES

[1] Karjalainen, M.A., "*A New Auditory Model for the Evaluation of Sound Quality of Audio Systems.*" Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Tampa, 4985.

[2] U.K. Laine, M. Karjalainen, and T. Altosaar, "*Warped Linear Prediction (WLP) in Speech and Audio Processing,*" Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Adelaide, 1994.

[3] J.P. Eatock and J.S. Mason, "A Quantitative Assessment of the Relative Speaker Discriminating Properties of Phonemes" Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, Adelaide, 1994.

[4] H. Hattori, "*Text-Independent Speaker Recognition using Neural Networks,*" Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing, San Francisco, 1992.