PERIODICA POLYTECHNICA SER. EL. ENG. VOL. 34, NO. 3, PP. 157-165 (1990)

ON SPEECH AND SPEAKER RECOGNITION USING NEURAL NET MODELS

J. S. MASON and E. C. ANDREWS

Department of Electrical and Electronic Engineering
University College, Swansea, UK

Received July 3, 1990; Revised January 11, 1991.

Abstract

The field of digital speech processing may be divided into three distinct and somewhat independent applications, namely speech recognition, speaker recognition and speech communications. Linear-predictive (LP) analysis techniques are used in all three areas to provide a compact signal representation that has a high information content.

This paper examines the somewhat conflicting tasks of speech and speaker recognition using perceptually based LP features. Using the recently developed multi-layer perceptron it is possible to construct a single architecture that may be trained to perform either one of these tasks using identical speech training data.

An Eset / 8 speaker and an alphabet / 26 speaker system are examined with both 5th and 14th order features. Both speech and speaker recognition tasks perform well confirming that the same structure fed with identical inputs can achieve both goals. It is found that the speaker specific information is contained predominantly in the higher order feature coefficients, with speech specific information concentrated in the lower order coefficients, confirming the results of Hermansky and Gu.

Keywords: linear-prediction, speech recognition, speaker recognition, neural networks, perceptual weightings.

Introduction

Automatic *speech* and *speaker* recognition systems have a lot in common: both have a front-end feature extraction process, followed by some form of pattern matching and decision making stages, and within each of these stages there are further similarities in the two applications.

Here we choose to concentrate on the feature extraction process. Short-term spectral estimates (MAKHOUL, 1975a; 1975b), especially in the form of cepstra derived from linear predictive (LP) analysis, are used widely for both *speech* and *speaker* recognition. Perhaps this is not surprising since the feature extraction process is designed to map the speech into a time series of convenient and representative patterns, recording time events in a form which provides data reduction and facilitates a meaningful distance measure.

The distance measure is especially important in the case of popular speech pattern matching processes based for example on dynamic time warping, and in this respect short-time spectral estimates are an obvious choice for both *speech* and *speaker* recognition. The cepstral representation is now extremely popular, largely due to the flexibility, convenience and performance derived from the associated Euclidean distance measure.

It is interesting to note that the same LP analysis approach now dominates speech coding research, where the primary objective is to maximise data reduction with no loss of intelligibility and with minimum loss of quality.

So the three major applications of computer speech processing, namely and in order of importance: speech coding, speech recognition and speaker recognition are all linked by LP analysis.

The use of LP models in coding is unquestionable proof that such features contain information components of both the text and the speaker. This raises the interesting question as to whether or not these two components embedded within the feature sequence can be somehow separated, since it is clear that the two recognition tasks require very different fundamental information. In fact the requirements are almost diametrically opposite, as depicted in *Fig. 1*.

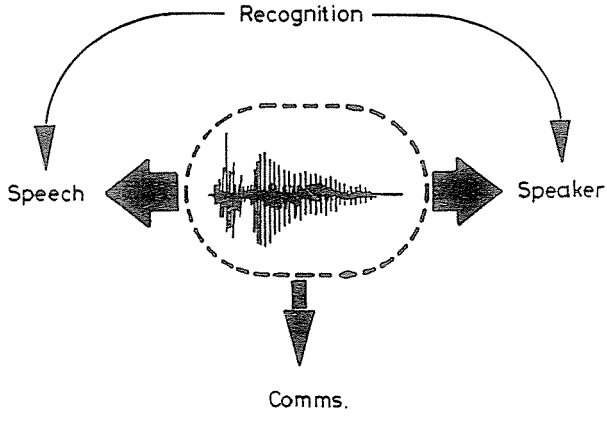

Fig. 1. Speech processing: LP analysis is common to all three applications even though the goals of each are very different.

Such an interpretation gains credence when dependence and independence of text and speakers are considered. In *speech* recognition, speaker-independence is a desirable characteristic and speaker-specific information is of significant nuisance value. Conversely in *speaker* recognition text-independence might be regarded as a desirable characteristic, in which

case the text information is a 'noise' to the pattern matching process. Thus for a given task, one component will act as a 'noise' to the other component, and *vice versa* for the other task.

One possible advantage of a neural net approach in this context is that the two information components might be separated or at least highlighted automatically. To this end we consider a given net structure first in a *speaker* recognition mode, then in a *speech* recognition mode. The net structure and its input remain unchanged; the *only* change is in the output labelling during training. We consider only one net structure, namely a multi-layer perceptron with a single hidden layer, trained using standard error–back propagation, gradient descent minimisation (RUMELHART, 1986), discussed below. The order of the feature analysis is also considered. The performance of nets using either PLP–5 or PLP–14 are compared in the two applications, confirming that the higher order coefficients contain primarily speaker specific information.

Speech Features and Perceptual Weightings

The front-end processing acts on the 'raw' speech samples, generating a sequence of feature vectors running along the time course of the utterance. Linear predictive analysis is now a *de facto* standard for speech processing.

If automatic *recognition* is our goal (speech or speaker), it might be useful to incorporate psychoacoustical knowledge into the analysis. One such successful approach, proposed originally by HERMANSKY et al (1985), is the perceptually-based linear prediction, called PLP. It is an all-pole representation after the 'auditory spectrum' has been mapped onto the speech, rather than an all-pole model of the 'raw' input speech itself as in the standard LP approach.

The *auditory* spectrum is derived by applying psychoacoustical transformations to the raw spectrum reflecting the

1. critical bands and masking effects,
2. equal loudness variations with frequency, and
3. intensity law.

These parameters can be modelled by:

1. $P_2(\omega) = \int_0^{\pi} C_k(\omega).P(\omega).d\omega$,
2. $P_3(\omega) = E(\omega).P_2(\omega)$ and
3. $Q_k(\omega) = {P_3}^{1/r}$,

where $P(\omega)$ is the power spectrum of the input speech, $C_k(\omega)$ is a set of critical band weighting functions, $E(\omega)$ is the equal loudness (pre-emphasis) curve, and $1/r$ is the intensity loudness conversion. The combination of

these effects gives the overall auditory spectrum model:

$$Q_k = \left\{ E(\omega) \int_0^{\pi} C_k(\omega).P(\omega).d\omega \right\}^{1/r} .$$

A number of researchers have used PLP, highlighting its merits in speaker-independent speech recognition. HANSON and WAKITA (1986) demonstrate the importance of an appropriate distance measure, the root-power-sum (RPS) being well-matched with PLP.

An important property in terms of computation is the relatively low order which gives good *speech* recognition performance. It is shown by HERMANSKY (1987) and GU and MASON (1988) that low order PLP, such as PLP–5 is particularly good for speaker-independent speech recognition, whereas higher orders, such as PLP–8 are better for speaker-dependent recognition. This implies that speaker-specific information is concentrated in the higher order coefficients of PLP. Hence for *speaker* recognition low orders of PLP would be expected to perform relatively poorly.

If the feature gives good performance in the speaker-independent mode, it is a reasonable hypothesis, following the above argument, that the speaker-specific information is less prevalent. Our experimental results here support this idea, and corroborate the findings of XU et al (1989) who highlight the relatively poor performance of PLP–5 in *speaker* recognition.

In the experiments described here two features are considered: PLP–5 and PLP–14. Both are computed from 25.6 ms windows (256 samples), overlapped by 50%. The vocabularies considered are the English Eset (b c d e g p t v) and the alphabet. Each utterance is linearly time normalised to 20 vectors to provide a constant length input to the net of either 100 (PLP–5) or 280 (PLP–14) coefficients.

The Multi-Layered Perceptron (MLP)

A single perceptron is a very simple device based loosely upon the structure of a neuron. A perceptron may have several inputs and outputs and may be trained by adjusting the relative importance of each input, or weight, together with its threshold.

The basic structure of an MLP is shown in *Fig. 2*. It is well-known that if only a single layer is used, without a so-called 'hidden layer', then the model can only divide the input space using hyperplanes. This means that only linearly separable problems may be solved using a single layer perceptron. A multi-layered structure, with one or more hidden layers, is not subject to this limitation.

Training is an important aspect of neural net work. The relatively recent development of the back-propagation algorithm (RUMELHART et al, 1986) gave the first practical solution to the problem of training multi-layer structures. The algorithm is an iterative one where a corpus of data is presented to the net and the outputs are compared with pre-labelled values to determine the errors. The net weights, represented by W, are updated and the process repeated, meaning that the task can be very computationally intensive.

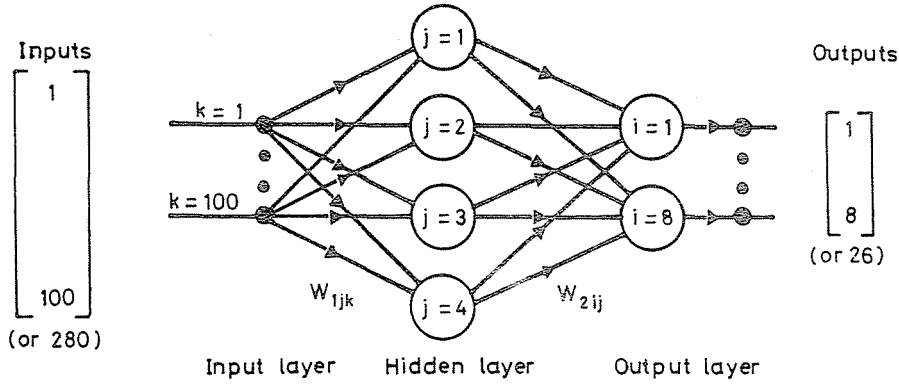

Fig. 2. The basic MLP structure

Performance of the net is measured during the training process by recording its status at intervals and testing the system, i.e. estimating its classification performance, using *unseen* data. Such results are averaged across speakers and utterances to provide performance profiles, examples of which are given below.

Experiments on Speech and Speaker Recognition

This group of experiments considers the two features, PLP–5 and PLP–14, applied to nets configured either for *speech* or for *speaker* recognition.

First a relatively small but difficult classification problem of distinguishing between a vocabulary of eight utterances spoken by eight people is considered. The task is difficult in that the chosen vocabulary is the Eset (b c d e g p t v); distinguishing between these utterances can be difficult even for humans.

The Eset contains 8 utterances, so for the *speech* recognition experiments the net must have 8 output nodes. To give the desired symmetry to

enable the identical net structure and inputs to be used subsequently for *speaker* recognition experiments, utterances from 8 people are used. Similarly 26 output nodes are required for the alphabet and hence utterances from 26 people are used to retain symmetry.

The net structure chosen has 40 nodes in the hidden layer. This is not claimed to be in any sense optimum, but merely a sensible practical value around which performance variations are found to be small. For PLP–5 and the Eset there are 100 (5 × 20) input nodes, together with the 40 hidden nodes and 8 output nodes, i.e. 100–40–8. For the PLP–14 Eset experiments the structure is 280–40–8, and in the case of the alphabet and PLP–14 the size is 280–40–26.

Two versions of each token from each of the speakers make up the training data; a third version is used for testing. This might be regarded as a relatively small amount of data to represent typical speech and speaker variations, but serves adequately here to illustrate the principles involved.

Eset / 8 speakers

The recognition profiles from this experiment are shown in *Fig. 3*.

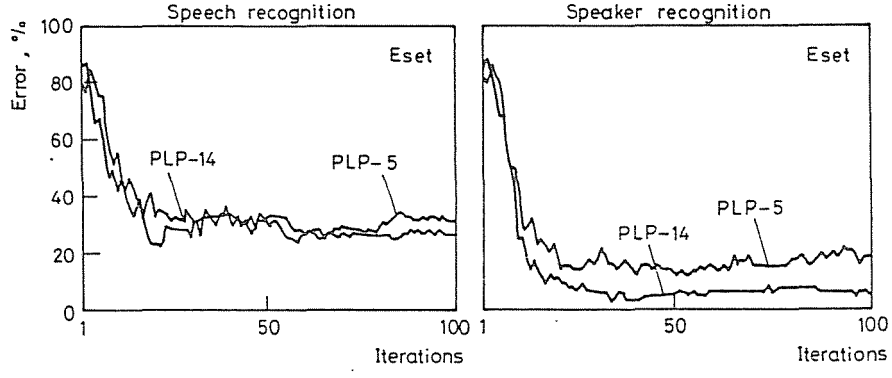

Fig. 3. Speech and speaker recognition performance from an identical net structure and training set, with 8 possible output classes. Note that performance improves with PLP–14 only in speaker recognition.

First consider the *speech* recognition results. There is no significant difference between the performance of the two features, PLP–5 and PLP–

14. Both achieve a minimum classification error of around 23%, which is a reasonable performance for this difficult vocabulary.

Next consider the *speaker* recognition performance. Using PLP–5 the minimum error is 13%, whereas using PLP–14 the error rate falls to under 4%. Clearly the extra information in PLP–14 is significant for speaker identification. It should be noted that the vocabulary chosen favours speaker recognition over speech recognition since the trailing 'e' sound contains little if any *speech* discriminating information.

Alphabet / 26 speakers

The performance of a new network trained either to distinguish between the 26 letters of the alphabet, or between 26 speakers, is shown in *Fig. 4*.

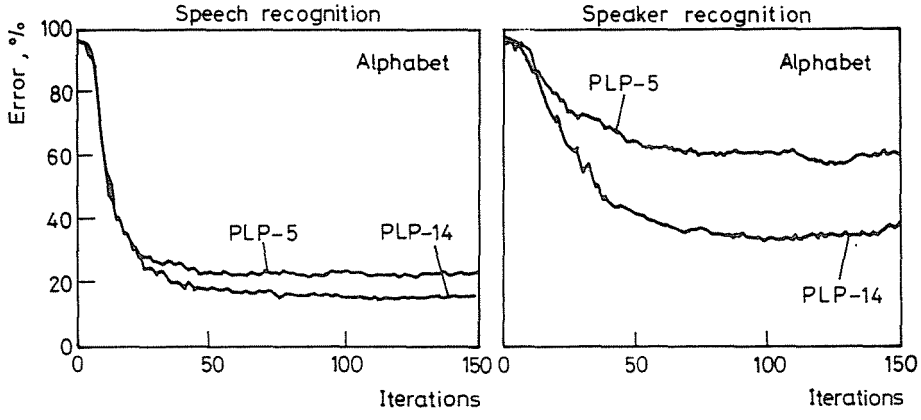

Fig. 4. Speech and speaker recognition performance from an identical net structure and training set, with 26 possible output classes. PLP–14 is superior to PLP–5 in both cases, but most significantly in speaker recognition where the error rate is halved.

The first point to note is the smoothness of the profiles compared to those from the 8×8 experiments in *Fig. 3*. This is due to the increased number of tests performed at each step: $26^2 = 676$ tests give a resolution of 0.2%, compared to $8^2 = 64$ tests and a resolution of 2% for the Eset.

For *speech* recognition PLP–5 has a minimum error of 22% compared to a PLP–14 performance of 15%. Thus in this larger problem there is some benefit in using the higher order feature.

The *speaker* recognition is now very much more text-independent compared with the Eset case, and clearly the task is that much more difficult. With PLP–5 the speaker recognition error does not fall below 60%, while PLP–14 gives a far better (though still not particularly good) performance of around 30% error. Better absolute performance would be expected with more training versions.

Conclusions

This paper examines speech and speaker recognition using a neural net approach. We have shown that an identical structure with identical inputs can be configured either for speech or for speaker recognition. This demonstrates the ease with which the same structure can be used for two different tasks.

The feature that we have used is perceptually weighted cepstra derived from linear prediction. We demonstrate clearly the importance of the higher orders for speaker recognition by comparing the performance of PLP–5 and PLP–14. The improvement derived from higher order features is far more noticeable for *speaker* recognition than it is for *speech* recognition. This is a first step towards separating speaker-specific and text-specific information, two components known to be embedded in the raw speech and in the extracted feature sequence. Current work is examining an integrated net structure with the goal of focusing on the knowledge representations within the one structure.

References

GU, Y. – MASON, J. S. (1988): A Speaker-Correlation Weighted RPS Distance Measure for Speech Recognition. *IEEE Symposium on Information Theory*, Kobe, Japan, June 1988, p. 72.

HANSON, B. A. – WAKITA, H. (1986): Spectral Slope Based Distortion Measures for All-pole Models of Speech. *Proc. ICASSP'86*, Tokyo, 7–11 April 1986, pp. 757–760.

HERMANSKY, H. – HANSON, B. A. – WAKITA, H. (1985): Perceptually Based Linear Predictive Analysis of Speech. *Proc. ICASSP'85*, Tampa, USA, 26–29 March 1985, pp. 509–512.

HERMANSKY, H. (1987): An Efficient Speaker-independent Automatic Speech Recognition by Simulation of Some Properties of Human Auditory Perception. *Proc. ICASSP'87*, New York, 6–9 April 1987, pp. 1159–1162.

MAKHOUL, J. (1975a): Spectral Linear Prediction. *IEEE Trans. ASSP-23*, pp. 283–296.

MAKHOUL, J. (1975b): Linear Prediction of Speech. *Proc. IEEE*, Vol. 63, pp. 561–580.

RUMELHART, D. E. – MCCELLAND, J. L. and the PDP Research Group (1986): *Parallel Distributed Processing*. Volume 1&2, MIT Press.

XU, L. – OGLESBY, J. – MASON, J. S. (1989): The Optimization of Perceptually-based Features for Speaker Identification. *Proc. ICASSP'89*, Glasgow, UK, 23–26 May 1989, pp. 520–523.

Address:

John S. MASON, Eddy C. ANDREWS
Department of Electrical & Electronic Engineering
University College of Swansea
Singleton Park, Swansea SA2 8PP
Wales, United Kingdom