Multi-source neural networks for speech recognition

Roberto GEMELLO, Dario ALBESANO, Franco MANA

CSELT - Centro Studi e Laboratori Telecomunicazioni
via G. Reiss Romoli, 274 - 10148 Torino - Italy
Tel.: +39-011-2286224 Fax: +39-11-2286207
email: gemello@cselt.it , albesano@cselt.it , mana@cselt.it

Abstract

In speech recognition the most diffused technology (Hidden Markov Models) is constrained by the condition of stochastic independence of its input features. That limits the simultaneous use of features derived from the speech signal with different processing algorithms. On the contrary Artificial Neural Networks (ANN) are capable of incorporating multiple heterogeneous input features, which do not need to be treated as independent, finding the optimal combination of these features for classification. The purpose of this work is the exploitation of this characteristic of ANN to improve the speech recognition accuracy through the combined use of input features coming from different sources (different feature extraction algorithms). In this work we integrate two input sources: the Mel based Cepstral Coefficients (MFCC) derived from FFT and the RASTA-PLP Cepstral Coefficients. The results show that this integration leads to an error reduction of 26% on a telephone quality test set.

1. Introduction

The two models that are presently obtaining the best performances in speech recognition are the standard Hidden Markov Models (HMM) and the connectionist derived hybrids that employ Neural Networks (NN) for the matching component (HMM-NN).

HMM are constrained by the condition of stocastic independence of their input features: this fact forces the use of a single input frame and limits the simultaneous use of features derived from the speech signal with different processing algorithms.
On the contrary, NN are capable of using multiple input features and finding optimal combination of these features for classification; these features do not need to be stochastically independent, and there is no need for strong assumpitons about the statistical distribution of input features, as required in HMM. This is a strength point of NN technology, and we intend to exploit it with multi-source neural networks.

In the last years the NN speech group at CSELT has developed and experimented a hybrid HMM-NN model [3][4]. After many studies on training strategies and network architecture, which brought to significative improvements, recently we have turned our attention on the possibility of integrating several input features to improve speech recognition accuracy. So we launched a research project about multi-source NN for speech recognition, with the aim of exploring several kinds of speech features and study their synergies as input of a HMM-NN model.
The first features we have planned to explore are, besides the standard Mel based Cepstral Coefficients (MFCC), some auditory inspired features, like RASTA-PLP and ear model derived features, and the Frequency Gravity Centers [6]. Then, we intend to study, among the others, Spectral Derivatives, Prosodic features and Subband Correlation features.

In this paper a first experimentation involving two features, MFCC and RASTA-PLP will be presented. After a description of the network architecture designed to integrate the two input features, some result will be presented, that show how the exploitation of synergy between different input features is a promising way to improve speech recognition accuracy.

2. Speech Modeling with Hybrid HMM-NN

Hybrid HMM-NN models integrate the ability of dealing with temporal patterns, typical of HMM, with the pattern classification power of NN. They inherit from HMM the modeling of words with left-to-right automata and the Viterbi decoding, delegating to a NN the computation of emission probabilities.

The recognition model we use is a hybrid HMM-NN model devoted to recognize sequential patterns [4]. Each class is described in terms of a left-to-right automaton (with self loops) as in HMM. The emission probabilities of the automata states are estimated by a Multi-layer Perceptron (MLP) neural network, instead than by mixtures of gaussians, while the transition probabilities are not

0-7803-5529-6/99/$10.00 ©1999 IEEE

considered. The MLP may be recurrent [3] or feedforward [4]: this architectural choice has to be decided experimentally case by case depending on the kind of acoustic units that are modeled. In the case of whole word models recurrent networks have proved to be superior while in the case of subword units, such as stationary-transitional units, feedforward MLP seems preferable.

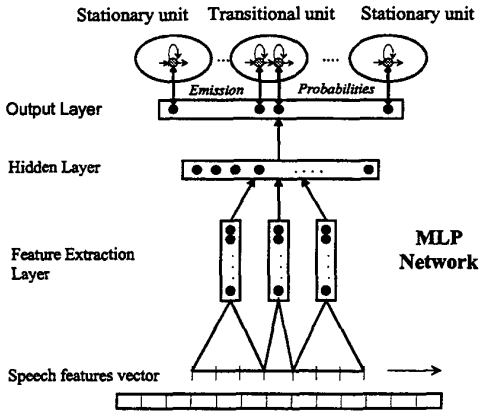

Fig. 1: Architecture of a Hybrid HMM-NN devoted to Stationary-Transitional Units Recognition

The MLP network has an input window that comprises some contiguous frames of the sequence, one or more hidden layers and an output layer where the activation of each unit estimates the probability P(Q|X) of the corresponding automaton state Q given the input window X.

A typical HMM-NN structure is depicted in fig. 1. The input window is 7 frames wide, and each frame contains the set of features extracted from the employed front-end. The first hidden layer is divided into three feature detector blocks, one for the central frame, and two for the left and right context. Each block is in its turn divided into six sub-blocks devoted to keep into account the six types of different input parameters. It was empirically found that this a priori structure is generally better than a fully connected layer. The second hidden layer is fully connected with the output layer that estimates the emission probabilities of the states of the word or phoneme automata, and is virtually divided in several parts, each one corresponding to an automaton.

The subword used for the acoustic modeling, first introduced by Fissore *et alii* [5], are called *Stationary-Transitional Units (STU);* they have very interesting features, and are very suitable to be modeled with neural networks.

These units are made up by stationary parts of the context independent phonemes plus all the admissible transitions between them. In the case of Italian language: there are 27 stationary parts and 348 transitions, for a total of 375 units. With these units a sequence of three phonemes *xpy* is modeled by the sequence of five units ...<*x*><*x-p*><*p*><*p-y*><*y*>... where <*x*>,<*p*>,<*y*> are the stationary parts of phonemes *x, p, y,* and <*x-p*>, <*p-y*> are the corresponding transitions between *x* and *p,* and *p* and *y*.
The background noise "@" is treated like a standard phoneme, so it is present its stationary part @ and its transitions to and from all the phonemes (e.g. @-a, ..., @-z, a-@, ..., z-@).
This set of STU is language dependent but domain independent, and represents a partition of the sounds of a tongue, like phonemes, but with more acoustic detail, including both the stationary parts and the transitions between them.

The main differences between our models and other hybrid HMM-NN models are:

- the different training procedure;
- the use of Stationary-Transitional units [4];
- the possibility to integrate different input features, described in this paper.

3. MFCC Features

The standard spectral features widely used in speech recognition are Mel Frequency Cepstral Coefficients (MFCC). Starting from digital signal a short-term Fourier analysis is performed each 10 ms extracting the power of the signal at each frequency. This powers are then band-grouped according to MEL spaced frequency subbands. The power of each band is then transformed into the logarithmic domain, so obtaining a Log-spectrum, which is, in its turn, transformed with a decorrelation transform named Discrete Cosine Transform (DCT) to obtain the final MFCC. The number of features extracted for each frame is 12 plus the total energy of each frame that is also retained. The first and second derivatives of cepstrals and energy are also computed as they provide important information about the dynamics of the signal.

Thus the standard features for speech recognition are 39 parameters for each 10 ms frame: 12 cepstrals, the total energy plus their first and second derivatives.

4. RASTA-PLP Features

Perceptual Linear Prediction is a feature extraction technique introduced by Hermansky and described in detail in [1]. In the PLP technique several well known properties of hearing are simulated by engineering approximations,

and in particular the nonlinear frequency scales (Bark, Mel), spectral amplitude compression, decreasing sensitivity of hearing at lower frequencies (equal-loudness curve) and large spectral integration .

In detail, first the short-term power spectrum is computed through FFT, and the critical band (Bark) nonlinear frequency resolution is emulated by integrating this spectrum under increasingly wider trapezoidal curves; then the bands are compressed with a logarithmic law and the unequal sensitivity of human hearing at different frequencies is emulated by an *equal loudness curve*; the nonlinear relation between the intensity of sound and its perceived loudness is simulated by an *intensity-loudness relation*, approximated by a cubic root compression. Finally, the all-pole model of the resulting spectrum is computed, according to LPC technique.

The RelAtive SpecTrAl (RASTA) technique was introduced by Hermansky and Morgan [2] as and engineering way to emulate the relative insensitivity of human hearing to slowly varying stimuli. It is a technique for dealing with slowly varying non-linguistic components of speech features due to convolutive noise (Log-RASTA) and to additive noise (J-RASTA). It consists in a band pass filtering performed on the logarithmic spectrum (Log-RASTA) or on the spectrum compressed by a ln(1+Jx) nonlinearity (J-RASTA). The key idea is to suppress constant factors in each spectral component of the short-term auditory-like spectrum prior the estimation of the all-pole model. The details about the RASTA-PLP technique can be found in [2]. The steps of RASTA-PLP processing are resumed in the following scheme:

Input: *Speech Signal*
1. Fast Fourier Transform
2. Critical-band Integration
3. Logarithm: ln(x) or ln(1+Jx)
4. Band-Pass Filtering
5. Equal-Loudness Curve
6. Power-Law of Hearing
7. Inverse Logarithm
8. Inverse Fourier Transform
9. LPC Analysis
10. Discrete Cosine Transform

Output: *Cepstral coefficients of RASTA-PLP analysis*

5. Network Structure

The network we use is designed to integrate in input the classical MFCC and the features extracted with RASTA-PLP. The network has an input layer, two hidden layers and an output layer.

the input layer is made up by 7 frames (one central frame plus a 3 frame left context and a 3 frame right context), each one composed by a block of MFCC and a block of RASTA-PLP features. The MFCC block contains: the Energy, 12 MFCC, their first derivatives and their second derivatives for a total of 13+13+13=39 values. The same structure takes place for the RASTA-PLP block.

The first hidden layer is divided into two blocks, one devoted to MFCC and the other to RASTA-PLP, each one of 315 units. Each of these blocks is divided, in its turn, into three blocks, one for the central frame, one for the left context and one for the right context. These blocks are made up by six sub-blocks devoted to the Energy, the Cepstrals, and their first and second derivatives, with a structure 5+30+5+30+5+30 = 105. These blocks are locally connected with the features which they are devoted to. Thus the first hidden layer performs a sort of local feature extraction.

The second hidden layer is composed by 250 units, fully connected with the previous layer. It performs an integration of the features extracted locally by the first hidden layer.

The output layer contains 379 units, one for each subword unit employed. While the hidden units are sigmoid units, the output are softmax, as they want to compute a probability distribution over the subword units.

6. Experimental Conditions

The training set used to train the HMM-NN model has the following characteristics:
- telephone quality, read speech;
- telephonic band 300-3400 Hz, sampled at 8KHz;
- 1136 speakers, evenly distributed among males and females;
- speakers from many Italian regions, with different accents;
- 4875 sentences and 3653 words phonetically balanced.

The test set is made up by 14473 utterances of isolated words pronounced by 1050 naive speakers from the list of all Italian city names (9329 words).
The MFCC were extracted with our standard front-end.
RASTA-PLP features were extracted by using the original software taken from ICSI. There are three principal types of processing available in the RASTA-PLP software: PLP, Log-RASTA and J-RASTA. PLP is the basic auditory inspired technique, Log-RASTA filtering was introduced to add robustness to convolutional noise, while J-RASTA attempts to handle both additive and convolutional noise. It is also possible to graduate RASTA processing by mixing its results linearly with pure PLP results.

7. Recognition Results

Several experiments were carried on to verify the improvements obtainable with the use of multiple input sources, and are summarized in table 1. First, the baseline results were obtained both with MFCC and with PLP, obtaining 83.45% and 84.07% respectively.

Table 1. Recognition Results

FRONT-END	WA%	Error Reduction
MFCC 12 cepstrals + E, d, dd	83.45	reference
PLP 12 cepstrals + E, d, dd	84.07	3.6%
Log-RASTA PLP 50% 12 cepstrals + E, d,dd	84.19	4.5%
MFCC + Log-RASTA PLP 50% 12 cepstrals + E, d,dd	87.59	25.0%
MFCC + J-RASTA PLP 50% 12 cepstrals + E, d, dd	87.83	26.4%

Then a set of explorative experiments, not reported here, were deployed to tune the optimal percentage of RASTA and PLP processing. A good proportion was 50% PLP and 50% RASTA, leading to 84.19%. This proportion was employed in the rest of experiments.

At this point, two multi-source experiment were conducted: the first one with MFCC plus Log-RASTA PLP, the second one with MFCC plus J-RASTA PLP. The first result was 87.59, with an error reduction of 25.0%, the second was 87.59, with an error reduction of 26.4%.
These results show that the conjunct and synergic use of different input sources is able to dramatically reduce the error rate.

8. Conclusions

The integration of multiple input sources is an important direction to improve hybrid HMM-NN speech recognition systems. In this paper we have proposed the integration of standard MFCC parameters with input features extracted with the RASTA-PLP processing.
The results obtained by the joint use of the two sets of features (about 26% error reduction) support our hypothesis that the synergy of different input features is useful for improving the recognition accuracy.

Future activities include the exploration of other kinds of features like, for example, spectral subband centroids, ear models, and prosodic features, with the aim of enriching the input to the neural network with different facets of the speech signal. Also the association of three or more different kind of features will be experimented.

Acknowledgments

The authors would like to thank Prof. Nelson Morgan of ICSI, Berkeley, that made available the RASTA-PLP software employed in this work.

References

[1] H. Hermansky, "Perceptual Linear Predictive Analysis of Speech", *J. Acoust. Soc. Am.*, pp. 1738-1752, 1990.

[2] H. Hermansky, N. Morgan, "RASTA Processing of Speech", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No 4, October 1994

[3] R. Gemello, D. Albesano, F. Mana, R. Cancelliere "Recurrent Network Automata for Speech Recognition: A Summary of Recent Work", in *Proc. of IEEE Neural Networks for Signal Processing Workshop*, Ermioni, Greece, September 1994.

[4] R. Gemello, D. Albesano, F. Mana "Continuous Speech Recognition with Neural Networks and Stationary-Transitional Acoustic Units", in *Proc. of IEEE International Conference on Neural Networks (ICNN-97)*, Houston, USA 1997.

[5] L. Fissore, F. Ravera, P. Laface, "Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition", in *Proc. of EUROSPEECH '95*, Madrid, Spain, September 1995.

[6] D. Albesano, R. De Mori, R. Gemello, F. Mana, "A Study on the Effect of Adding New Dimensions to Trajectories in the Acoustic Space", in *Proc. of Eurospeech'99*, Budapest, Hungary, to appear.