Politecnico di Torino

Porto Institutional Repository

[Proceeding] Synergy of Spectral and Perceptual Features in Multi-source Connectionist Speech Recognition

(Article begins on next page)
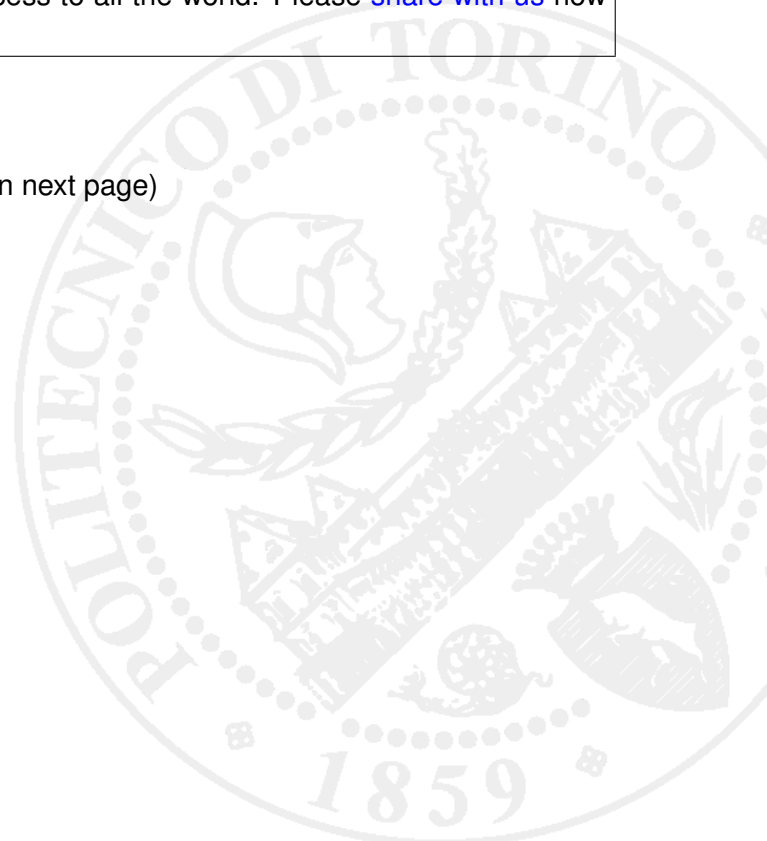
SYNERGY OF SPECTRAL AND PERCEPTUAL FEATURES
IN MULTI-SOURCE CONNECTIONIST SPEECH RECOGNITION

Roberto Gemello *, Loreta Moisa +, Pietro Laface +

* CSELT - Centro Studi e Laboratori Telecomunicazioni
Via G. Reiss Romoli, 274 - 10148 Torino - Italy
Tel.: +39-011-2286224   Fax: +39-11-2286207
+ Politecnico di Torino – Dipartimento di Automatica e Informatica
C.so Duca degli Abruzzi, 24 – 10129 Torino - Italy
email: gemello@cselt.it , moisa@cselt.it, laface@polito.it

ABSTRACT

The combined use of different set of features extracted from the speech signal with different processing algorithms is a promising approach to improve speech recognition performances.

Artificial Neural Networks are well suited to this task since they are able to use directly multiple heterogeneous input features to estimate a near optimal combination of them for classification, without being constrained by a priori assumptions on the stochastic independence of the input sources.

This work shows how we have taken advantage of these characteristics of Neural Networks to improve the recognition accuracy of our systems. In particular, three set of input features have been considered as sources in this work: Mel based Cepstral Coefficients derived from the FFT spectrum, RASTA-PLP Cepstral Coefficients, and a set of features that describe the dynamics of the FFT power spectrum along the frequency dimension, instead of the usual time dimension.

The experimental results confirm the usefulness of the proposed approach of feature integration that leads to a significant error reduction both on isolated and continuous speech recognition tasks on a large telephone speech test set.

1. INTRODUCTION

The investigation of acoustic features that can capture different facets of the speech signal is an important issue in speech recognition research because the recognizer performance is strongly affected by the accuracy of the features and by their robustness to noise. Several features have been proposed in the literature, but two of them are so widely used to be considered as "standard" acoustic front-ends: Mel Cepstral Coefficients (MFCC) [2] , and cepstral parameters derived from a Perceptual Linear Predictive analysis PLP [9] .

In our experience, the recognition performance obtained by the individual use of these features is comparable, but we observed that the resulting errors are often different and, to some extent, not correlated. We decided, therefore, to explore the benefits of the parallel use of different and partially complementary features to enrich the input information available to the pattern matching module, or at least to feed this module with different facets of the same information.

It is worth noting that an effective exploitation of additional features is strongly affected by the technique employed for the acoustic matching. The acoustic models that are most popular in speech recognition are Hidden Markov Models (HMM). Standard HMMs, using diagonal covariance continue density mixtures, are constrained by the condition of stocastic independence of their input features. This condition limits the simultaneous use of different set of features derived from the speech signal because it cannot be realistically assumed that they are independent. Another deficiency of HMMs is that the observation vector is limited to one frame. The first and second derivatives of the parameters are typically included in the observation vector to partially account for a larger temporal context.

Feature transformations obtained through Linear Discriminant Analysis (LDA) have been proposed to overcome this limitations, and a join use of different features (MFCC and -PLP) has delivered small improvements in the framework of HMM [8].

An alternative approach to continuous density HMM modeling is based on connectionist hybrids that employ Neural Networks (NN) for the matching component (HMM-NN). Neural Networks are able to combine several heterogeneous input features to find a locally optimal solution to their classification task. The NN input features do not have to be stochastically independent, neither there is the need for strong assumptions about their statistical distribution, as is required for HMMs. This allows NNs to exploit both multiple input sources and large frame contexts.

In the last years we have developed and experimented hybrid HMM-NN models [4][5] obtaining results that are comparable with those delivered by the HMM technology. Then, several sets of features have been explored with the aim of studying their

synergy with the MFCC features when they are supplied as input to a multi-source HMM-NN. In particular, results have been presented for combination of MFCC parameters with Frequency Gravity Centers [1], Ear Model derived features [6], and also preliminary results for combinations of MFCC with RASTA-PLP [7] parameters.

In this paper an experimentation involving three sets of features: MFCC, RASTA-PLP [10] and Spectral Frequency Derivatives (SFD) [11][12] is presented. After a description of the network architecture designed to integrate the input features, the results on two test corpora will be presented, showing that, in a connectionist framework, the exploitation of the synergy among different input features significantly improves the recognition accuracy .

2. SPEECH MODELING WITH HYBRID HMM-NN

Hybrid HMM-NN models integrate the ability of dealing with temporal patterns, typical of HMMs, with the discriminant classification capabilities of NNs. They share with HMMs the sub-word model topology described by left-to-right automata and the Viterbi decoding algorithm, but the emission probabilities of the states of the models are computed by a properly trained NN, rather than derived from their associated probability distribution modeled by mixtures of Gaussians.

In particular, we use a hybrid HMM-NN model where each sub-word 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, while the transition probabilities are not reestimated. We have experimented both with recurrent [4] or feedforward [5] MLP architectures. Recurrent networks have proved to be superior in whole word model training, while feedforward MLP are preferable for training sub-word units.

We refer to the sub-word used for the acoustic modeling as Stationary-Transition Units (STU) [3]. The set of these units include the stationary parts of the context independent phonemes (less affected by the phonetic context) and all the admissible transitions between them. In the case of the Italian language: we defined 27 stationary units and 348 transition units, for a total of 375 units. With this definition of 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 the 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 handled as a phoneme, so we model its stationary part @ and its transitions to and from all the phonemes (e.g. @-a, ..., @-z, a-@, ..., z-@).

A stationary unit corresponds to one output unit of the NN, while transition units correspond to two network output units.

The set of STU is language dependent, but domain independent, it seems to give a good tradeoff between the accuracy and trainability of context dependent models, even compared with set of tied triphones obtained by means of classification trees. This modeling has been applied with good results to other languages like English, Spanish and German.

3. FEATURES

The experiments presented in Section 6 refer to three set of features: MFCC, RASTA-PLP and Spectral Frequency Derivatives. Our settings for these features are described in the following Subsections.

3.1 MFCC

MFCC are one of the most popular set of spectral features used in speech recognition. Our systems use a typical configuration that includes 39 parameters for each frame of 10 ms: 12 cepstral coefficients (C0 is discarded) and the total energy, plus their first and second derivatives.

3.2 RASTA-PLP

Perceptual Linear Prediction is a feature extraction technique introduced in [9] that obtains a smoothed spectrum by fitting the parameter of an all-pole model spectrum to a Mel-scale spectrum.

The PLP features are often further processed by means of the RelAtive SpecTrAl (RASTA) technique introduced in [10] as an engineering way to emulate the relative insensitivity of human hearing to slowly varying stimuli. The central 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. Again 12 RASTA-PLP parameters are extracted every 10 ms together with the energy, and the observation frame includes their fist and second derivatives.

3.3 SPECTRAL FREQUENCY DERIVATIVES

Studies on the auditory system show that the ear is sensible to variations in frequency of the signal. The use of features related to the frequency variations of the signal has been proposed in [11] and [12].

The power spectrum $S(t, f)$ is generally obtained by the FFT of the signal and it can be referred to as:

$$S(t, f) = \left\{ S(t_1, f), \ldots, S(t_i, f), \ldots, S(t_N, f) \right\}$$

where: $S(t_i, f)$ is a vector that represents the energy distribution of the signal along the frequency domain (128 elements in our case) at time t_i, and N is the duration of the utterance. This vector is normalised using a log function, chosen according to the perceptual theory:

$$S'(t_i, f) = \log(1 + \alpha \cdot S(t_i, f))$$

(In our experiments we set $\alpha = 1$)

The features that we considered in this work are the first and second spectral frequency derivatives computed along the 128-vector $S'(t_i, f)$. The result of the first derivative $\partial f(S'(t_i, f))$ is a 128-vector of derivative values that, we further group in 13 sub-bands according to the Mel scale in order to obtain a number of parameters comparable with the other

feature. The second derivative $\partial^2 f(S'(t_i, f))$ is computed from the 128-vector of the first derivatives.

4. NETWORK STRUCTURE FOR MULTI-SOURCE FEATURES

The integration of multi-source features is rather straightforward in our network whose skeleton is given in Fig. 1 for the combination of MFCC and RASTA-PLP parameters. The combination with the SFD features is obtained with the same architecture.

The MLP network has an input window that includes a set of contiguous frames of the observation sequence, two hidden layers, and an output layer where the activation function of each network unit estimates the probability P(S|O) of the corresponding state S given the input window O.

The input window has a width of 7 frames, (one central frame, 3 frames for the left context, and 3 frames for the right one). One superframe includes the set of MFCC and the other one the set of RASTA-PLP parameters described in Section 3.

The first hidden layer is divided into two blocks, devoted to the MFCC and to the RASTA-PLP parameters respectively. Each of these blocks is divided, in its turn, into three feature detector blocks, one for the central frame, one for the left context and one for the right context. Every "context" block is sub-divided into 6 blocks devoted to the processing of homogeneous input parameters: the Energy, the MFCC (RASTA_PLP) coefficients, and their first and second derivatives. It has been found that this structure is less expensive and generally better than a fully connected layer.

Every energy unit is connected to 5 units of the first hidden layer: this allows the network to perform a kind of soft vector quantization of the energy values. The 12 spectral parameters of the input layer are fully connected to 30 units in the first hidden layer allowing a first nonlinear combination. The same structure is used for the derivatives. Thus, the number of hidden units connected to the units of the input layer fed by the central frame is 105 ((30+5) * 3 considering the derivatives), and the total number of hidden units of the MFCC extraction block is 315 considering the central, left and right context. Using this architecture, the first hidden layer performs an implicit local feature extraction by a nonlinear combination of the input parameter with a context of 7 frames.

The connection structure is repeated for the RASTA-PLP parameters, the two local feature extraction layers are independent but fully connected with the second hidden layer that comprises 300 units and performs a global integration combining the set of heterogeneous features into a new set of features.

The output of this layer is fully connected with the output layer. The latter contains 686 units, one for each stationary unit and two for each transition unit. While the hidden units are sigmoidal units, the output units use a softmax function, in order to compute a probability distribution over the sub-word units.

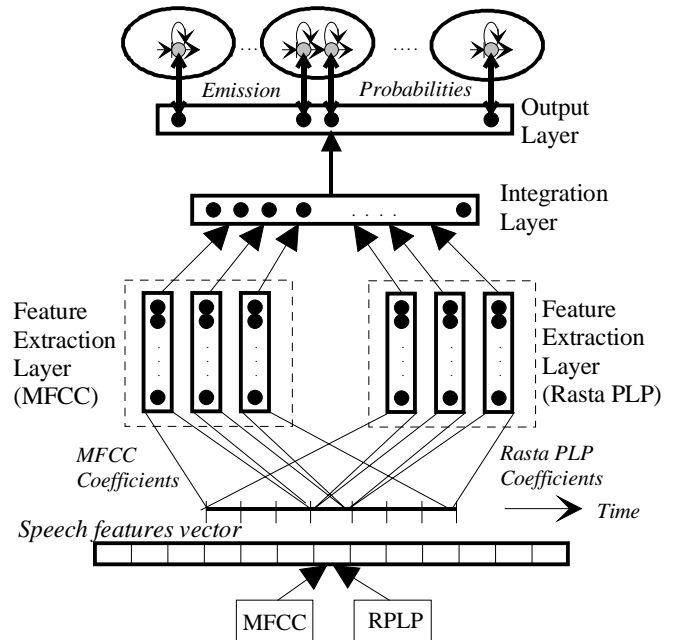The NN training procedure is described in detail in [5].

Figure 1. Architecture of a Hybrid HMM-NN with MFCC and RASTA-PLP input features.

5. EXPERIMENTAL CONDITIONS

The speech databases used to train the HMM-NN model have the following characteristics: telephone quality, read speech, bandpass filtered between 300-3400 Hz, sampled at 8KHz. About 6000 speakers, evenly distributed among males and females, from many Italian regions, with different accents, contributed to the database collection. The database includes 30 hours of labeled speech with about 9000 sentences and 19000 words phonetically balanced.

6. RECOGNITION RESULTS

The performance of the multi-source systems were tested both with isolated words and with continuous speech. The first test, with a vocabulary size of 9329 words, includes 14473 utterances of isolated words pronounced by 1050 naive speakers from a list of Italian city names. The second one, with a vocabulary of 9400

FRONT-END	WA	Error Reduction
MFCC 12 cepstral + E, d, dd	**89.29**	B_{MFCC}
MFCC + SFD	**91.14**	17.27%
RPLP 12 cepstral + E, d,dd	**89.79**	B_{RPLP}
MFCC + RASTA-PLP 12 cepstral + E, d, dd	**91.61**	21.6%

Table 1. Recognition results for isolated words

words, includes 4296 phonetically balanced sentences of read telephone speech with an average length of 6 words.

Four HMM-NN models were trained, with different input features configurations, and compared on the two test database. A first set of recognition experiments, carried out on the isolated word test set, is summarized in Table 1. A second set of experiments was carried out on the continuous speech test set: it is summarized in Table 2. No language modeling has been used in these tests to account for the contribution of the acoustic features only.

The baseline result of 89.29% and 58.1 (B_{MFCC}) were obtained using the MFCC features alone for the isolated word and continuous speech test respectively.

Then a NN combining MFCC and SFD features was trained. The additional computation for obtaining the frequency derivative feature is negligible. An improvement has been obtained for the isolated words, but it was disconfirmed by the disappointing result (3.3% of error increase) obtained on continuous speech (see Table 2).

It was decided, thus, to consider the possible contributions of the RASTA-PLP features.

The results for the baseline RASTA-PLP network are 89.79%. and 60.0% respectively (B_{RPLP} in Table 1 and 2). Please notice that these values are comparable with those obtained by the MFCC NN.

Then the multi-source experiment was conducted combining the RASTA-PLP features to the MFCC ones. The result was 91.61% and 65.5% respectively, with significant error reductions of 21.6% (17.7%) with respect to the B_{MFCC} results.

FRONT-END	WA	WI	WD	WS	E.R.
MFCC 12 ceps + E, d, dd	**58.1**	3.7	9.2	29.0	B_{MFCC}
MFCC + SFD	**56.7**	8.6	8.3	26.5	-3.3%
RPLP 12 cep + E, d, dd	**60.0**	3.4	8.3	28.3	B_{RPLP}
MFCC + RPLP 12 cep + E, d, dd	**65.5**	3.5	7.4	23.6	17.7%

Table 2. Recognition results for continuous speech

These results show that the use of different input sources inside the connectionist recognition framework is able to significantly reduce the error rate. Other experiments, not reported here, show similar improvements also on other test cases, in particular in a keyword-spotting case and on real data from an on-field application about railway timetable queries.

7. CONCLUSIONS

In this paper we have proposed a NN architecture for the integration of standard MFCC parameters with spectral frequency derivative features and with RASTA-PLP parameters.

The results obtained by the joint use of the MFCC and RASTA-PLP features support our hypothesis that the synergy of different input features is useful for improving the recognition accuracy.

The average computational overhead due to the double front-end can be estimated in a 20%, and it is fully supported by conventional hardware, allowing the recognizer to run in a real time on a Pentium II PC.

8. REFERENCES

[1] 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, 1999, pp.1503-1506.

[2] S.B. Davis, P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", in *IEEE Transactions on Acoustic Speech and Signal Processing,* Vol. ASSP-28, pp.357-366, Aug. 1980.

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

[4] 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.

[5] 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, pp.2107-2111.

[6] R. Gemello, D. Albesano, F. Mana, "Synergy of Spectral and Ear Model Features for Neural Speech Recognition", in *Proc. of International Conference on Artificial Neural Networks - ICANN '99,* Edimburgh, Scotland, September 1999.

[7] R. Gemello, D. Albesano, F. Mana, "Multi-source neural networks for speech recognition", in *Proc. of International Joint Conference on Neural Networks (IJCNN'99),* Washington, July 1999.

[8] R. Haeb-Umbach, M. Loog, "An Investigation of Cepstral Parameterisation for Large Vocabulary Speech Recognition", in *Proc. of Eurospeech'99,* Budapest, Hungary, 1999, pp.1323-1326.

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

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

[11] T. Nitta, "*A Novel Feature-Extraction for Speech Recognition Based on Multiple Acoustic-Feature Planes*", Proceedings of ICCASP-98.

[12] R. Oka , H. Matsumura, "*Speaker Independent Word Speech Recognition using the blurred orientation pattern obtained from the vector field of spectrum*", Proceedings of IJCPR-88.