Loquendo
VOCAL TECHNOLOGY AND SERVICES

Hybrid HMM/Neural Network based Speech Recognition in Loquendo ASR

Roberto Gemello, Franco Mana, Dario Albesano
Loquendo

roberto.gemello@loquendo.com

Abstract

This paper describes hybrid Hidden Markov Models / Artificial Neural Networks (HMM/ANN) models devoted to speech recognition, and in particular Loquendo HMM/ANN, that is the core of Loquendo ASR.
While Hidden Markov Models (HMM) is a dominant approach in most state-of-the-art speaker-independent, continuous speech recognition systems (and commercial products), Artificial Neural Networks (ANN) are universally known as one the most powerful non-linear methods for pattern recognition, time series prediction, optimization and forecasting.
Hybrid HMM/ANN, introduced in the nineties for speech recognition, is presently a very competitive alternative to HMM, both in terms of performances and recognition accuracy.
HMM/ANN combines the advantages of both approaches by using an ANN (a multilayer perceptron) to estimate the state dependent observation probabilities of a HMM, instead of Gaussian mixtures, while the temporal aspects of speech are dealt with by left-to-right HMM models.
HMM/ANN can provide discriminative training, are capable of incorporating multiple input sources, and have a flexible architecture which can easily accommodate contextual inputs and feedbacks.
Furthermore, ANN are typically highly parallel and regular structures, which makes them especially suited for high-performance architectures and optimized implementations.

1. Introduction

An artificial neural network (ANN) is an interconnected group of artificial neurons that uses a computational model for information processing based on a connectionist approach. ANN are widely used in many fields of engineering, in particular for pattern recognition [1], classification and prediction.

Although Artificial Neural Networks have been shown to be quite powerful in static pattern classification, their formalism is not very well suited to addressing automatic speech recognition (ASR). In fact in ASR there is a time dimension which is highly variable and difficult to handle directly in ANN. The ASR problem can be stated as follows: how can an input sequence (e.g., a sequence of spectra or spectral derived coefficients) be properly classified into an output sequence (e.g., sequence of phonemes, words or sentences) when the two sequences are not synchronous, since there usually are multiple inputs associated with each phoneme or word?

Several neural network architectures have been developed for (time) sequence classification, including:

- Static networks with an input buffer to transform a temporal pattern into a spatial pattern [2][5]
- Recurrent networks that accept input vectors sequentially and use a recurrent internal state that is a function of the current input and the previous internal state [3][4].
- Time-delay neural networks, approximating recurrent networks with feed-forward networks [6].

In the case of ASR, all of these models have been shown to yield good performance (sometimes better than HMM) on short isolated speech units. By their recurrent aspect and their implicit or explicit temporal memory they can perform some kind of integration over time. However, neural networks by themselves have not been shown to be effective for large scale recognition of continuous speech. To overcome this problem some authors introduced, in the early nineties, a new approach that combine ANN and HMM for large vocabulary continuous speech recognition [5].

2. HMM/ANN Hybrid Systems

Most speech recognition systems use continuous density HMM as a standard approach.

Some years ago, a new formalism particularly well suited to sequential patterns (like speech and handwritten text) and which combines the respective properties of ANN and HMM was proposed and successfully used for difficult ASR (continuous speech recognition) tasks [5]. This system, usually referred to as the hybrid HMM/ANN, combines HMM sequential modeling structure with ANN pattern classification.

As in standard HMM, hybrid HMM/ANN systems applied to ASR use a Markov process to temporally model the speech signal. The connectionist structure is used to model the local feature vector conditioned on the Markov process. This hybrid is based on the theory that ANN can estimate class (posterior) probabilities for input patterns. This probability can then be used, after some modifications as local probabilities in HMM. In practice the hybrid HMM/ANN system replaces the Gaussian mixture HMM state-dependent observation probability with estimates computed by MLP, keeping the HMM topology unchanged.

Advantages of the HMM/ANN hybrid for speech recognition include:

- a natural structure for discriminative training: in fact MLP neural networks are trained in a discriminative way with error-back propagation ;
- no strong assumptions about the statistical distribution of the acoustic space: this is a theoretical property of ANN, unlike standard HMM which assumes that all the subsequent input frames are independent, that in speech is clearly not realistic;
- parsimonious use of parameters: the use of a distributed model like ANN allows good results to be obtained with a reduced number of parameters
- better robustness to insufficient training data: ANN are recognized as being very good at generalizing;
- an ability to model acoustic correlation (using contextual inputs or recurrence): unlike standard HMM, HMM/ANN can use a temporal context extensible as needed.

In recent years these hybrid approaches have been compared with the best classical HMM approaches on a number of ASR tasks. In cases where the comparison was controlled (e.g., where the same system was used in both cases except for the means of estimating

emission probabilities), the hybrid approach performed better when the number of parameters were comparable, and about the same for some cases in which the classical system used many more parameters. Also, the hybrid system was quite efficient in terms of CPU and memory run-time requirements.

More generally, though, complete systems achieve their performance through detailed design, and comparisons are not predictable on the basis of the choice of the emission probability estimation algorithm alone.

3. Loquendo Hybrid HMM/ANN ASR

3.1 Basic model

Hybrid HMM/ANN models combine the ability to deal 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.

Loquendo (and previously CSELT) has been working on NN for speech recognition since 1988, and developed its hybrid HMM/ANN model in the early nineties, evolving it year by year. Starting from a first model dealing only with isolated words [7], the extension to open vocabulary [8] by employing a proprietary set of subwords was developed. Then, the use of a patented ANN acceleration technique and ad-hoc optimizations for the run-time module [9] greatly reduced the computational effort required, allowing large networks to run in real time on standard PCs. Recently the study and experimentation of new techniques for ANN adaptation has opened the door to important applications in cases where the adaptation to speaker or environment is necessary [10]. Presently Loquendo hybrid HMM/ANN recognition technology is the core of Loquendo ASR, used worldwide in several successful applications.

The main differences the between Loquendo model (also referred to below as NNA – *Neural Network Automata*) and other hybrid HMM-NN models are:

- a time/feature locally connected architecture in the first hidden layer;
- the use of Stationary-Transitional units (described in section 3.2);
- a different training procedure [8];
- an efficient way to extend the model to deal with Multi-Source input (described in section 3.2);
- the use (at run time) of patented acceleration methods and optimized assembly code;
- A new original adaptation method [10].

Loquendo hybrid HMM/ANN is a model devoted to recognizing sequential patterns, named *Neural Network Automata (NNA)*. 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, rather than by Gaussian mixtures, while the transition probabilities are not considered. The MLP may be recurrent or feedforward: 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 phonemes or STU acoustical models, feedforward MLP seems preferable.

The NNA has an input window that comprises some contiguous frames of the sequence to incorporate contextual information, 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.
An NNA has many degrees of freedom: the architecture of the MLP, the input window width, the number of automaton states for the different words or phonemes employed. A general NNA structure for the Basic and Multi-Source streaming, where additional and alternative features can be added to the basic one, is depicted in figure 1. Focusing on the basic architecture, the input window is 7 frames wide, and each frame contains 39 parameters (log Energy, 12 MFCC (or RPLP), and their first and second derivatives). 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 take 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 into several parts, each one corresponding to an automaton.

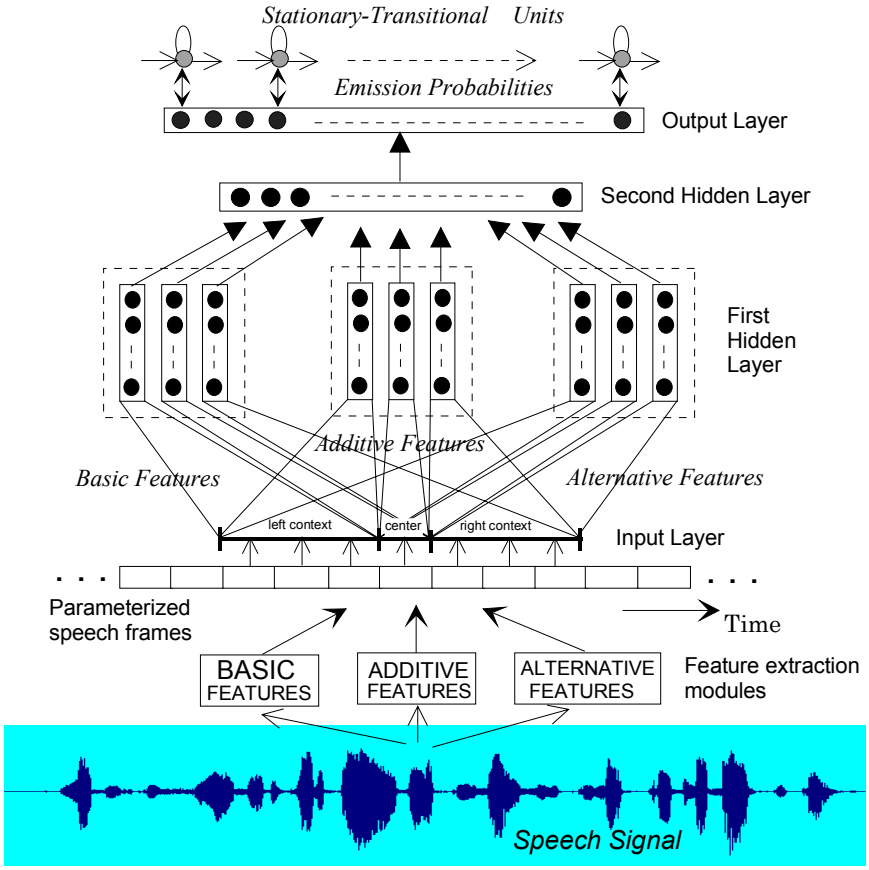

Figure 1. General Architecture of Loquendo Hybrid HMM/ANN model

3.2 Stationary-Transitional Acoustic Modeling

Neural networks are discriminative classifiers and are well suited for output classes which are a partition of the output space, that is, that are not overlapping and cover the whole output space. For this reason HMM/ANN hybrids usually model context independent phonemes (which are a partition of the sounds) and not context-dependent biphones and triphones (which are overlapping).

An original way of modeling contextual information, very much suited to be used with ANN, was introduced by Loquendo researchers in [11]. The proposed units, called *Stationary-Transitional Units (STU)*, are made up by stationary parts of the context independent phonemes plus all the admissible transitions between them. 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 (both phonemes and diphones). The modeling of STU with NNA can be realized by assigning one output unit of the MLP to each stationary unit, and one or two states to each transitional unit.

The use of STU greatly improves the recognition accuracy w.r.t. context independent phonemes, while maintaining the unit number contained w.r.t. triphones. STU are a distinctive feature of Loquendo ASR.

3.3 Multi-Source Input

The use of multiple knowledge sources can be very useful to improve recognition performances. The Loquendo hybrid HMM/ANN model is able to utilize two (or more) different sets of input features (see Figure 1). In the case of Multi-Source input, the speech signal is analyzed by several signal-processing algorithms to extract several kinds of features: the *basic features* (the standard MFCC or RPLP coefficients), the *additive features* (that is, those features that are not independent but should be used in addition to the basic ones, like Pitch, periodicity, gravity centers, formants, signal-noise ratio) and the *alternative features* (that is those features that are sufficiently informative and could be used instead of the basic features, e.g. Wavelet derived parameters, ear-model features). Till now, we have investigated successfully the use of RASTA-PLP, MFCC, Wavelet transform, Pitch and ear-model, in various combinations. All the used features are assembled into each input frame. Then the input frames are loaded into the network.

The input layer is made up, as before, of 7 frames (one central frame plus a 3 frame left context and a 3 frame right context), each one composed by a block of basic features, one or more blocks of additive features (if present) and one block of alternative features (if present). The hidden layers of the network are structured to separately manage the different sources in the lower layer and then to integrate them in the higher layers.

The problem of integrating Multi-Source information into the computation of the emission probability modeled by the NNA is delegated to the neural network capability to deal with a generic input data vector, without any statistical assumption of independence of the elements.

The use of multi-source input, e.g. RPLP+MFCC or RPLP+Wavelets can lead to an error reduction of up to 15-20% on large vocabulary speech recognition tasks.

3.4 Adaptation

The use of general acoustic models in the Loquendo ASR Automatic Speech Recognition engine provides excellent performance in a vast array of applicative conditions. However, it is important to be able to adapt the acoustic models in such a way as to take into account acoustic material acquired in the field. For instance, adapting recognition to a specific individual's voice, or to a set of individuals, or to the telephone environment at large.

Alongside the new version of Loquendo ASR, Loquendo recently released an innovative adaptation technique, called the Loquendo Acoustic Model Adaptation Tool, which is based on a patented technology.

Unlike standard HMM, where many adaptation algorithms are available (MAP, MLLR), there is limited available literature on proposals for the adaptation of HMM-NN. A standard technique is known as Linear Input Network (LIN) and was published in the nineties [12]. A LIN is a single layer network that performs a linear mapping of the space of input parameters. This network is placed before the HMM-NN, already trained in a speaker independent manner, and is trained with the adaptation material.

LIN was the first adaptation technique to be implemented in Loquendo AMA. Following extensive experimentation, this technique was judged to be valid but not completely satisfactory. Thus, Loquendo developed a brand new technique named LHN, which was patented in June 2005 and presented in [10]. LHN is a linear transform applied not on the input space but on the hidden unit activations space. The motivation is that the activations of an internal layer represent a projection of the input pattern into a space where it should be easier to learn the classification or transformation expected at the output of the network. As for the LIN, the values of an identity matrix are used to initialize the weights of the LHN. The weights are trained using a standard back-propagation algorithm keeping frozen the weights of the original network. It is worth noting that, since the LHN performs a linear transformation, once the adaptation process is completed, the LHN can be removed combining LHN weights with the ones of the next layer using simple matrix operations [10].

A new solution, called Conservative Training, is also proposed, that compensates for the lack of adaptation samples in certain classes.

Supervised adaptation experiments with different corpora and for different adaptation types have been performed. The results show that the proposed approach always outperforms the use of transformations in the feature space, and yields even better results when combined with linear input transformations.

4. Discussion

Bourlard and Morgan, in their seminal work on HMM/ANN [5] reported several valid reasons to adopt HMM/ANN in a speech recognition system instead of the classic HMM:

- ANN can provide discriminant-based learning; that is, models are trained to minimize the error rate while maximizing the distance between the correct model and its rivals.
- ANN can generate, in theory, any kind of non-linear functions of the input [2].
- Because ANN are capable of incorporating multiple constraints and finding optimal combinations of constraints for classification, there is no need for strong assumptions about the statistical distributions of the input features (while standard HMM requires statistical independence).
- ANNs have a flexible architecture which can easily accommodate contextual inputs and feedbacks.
- ANNs are typically highly parallel and regular structures, which makes them especially amenable to high–performance architectures and implementations.

From our experience, we can report that HMM/ANN usually outperforms classic HMM in terms of recognition performances, and is more accurate in performing phonetical decoding, which can be a building block for many voice recognition tasks.
Furthermore, the parallel structure of ANN has allowed us to obtain high run-time performances, thanks to the joint use of patented acceleration methods, optimized linear algebra libraries and assembly optimizations.
A weak point still present in HMM/ANN, i.e. their inferior adaptation capabilities w.r.t. classical HMM, has recently been overcome by new methods recently patented and released by Loquendo.

References

[1] Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*, Oxford: Oxford University Press.
[2] R. P. Lippmann. Review of neural networks for speech recognition. *Neural Computation*, 1(1):1--38, 1989.
[3] G. Kuhn, R. L. Watrous, and D. Ladendorf. Connected recognition with a recurrent network. *Speech Communication*, 9(1):41--48, 1990.
[4] T. Robinson and F. Fallside. A recurrent error propagation network speech recognition system. *Computer Speech and Language*, 5:259--274, 1991.
[5] H. Bourlard and N. Morgan. *Connectionist Speech Recognition---A Hybrid Approach*. Kluwer Academic, 1993.
[6] K. J. Lang, A. H. Waibel, and G. E. Hinton. A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3(1):23--43, 1990.
[7] D. Albesano, R. Gemello and F. Mana, "Word Recognition with Recurrent Network Automata", in Proc. IJCNN 92, Baltimore, June 1992, pp. 308-313.
[8] R. Gemello, D. Albesano, F. Mana, "Continuous Speech Recognition with Neural Networks and Stationary-Transitional Acoustic Units", in Proc. of IEEE Conference on Neural Networks (ICNN-97), Houston, USA 1997.
[9] D. Albesano, F. Mana, R. Gemello, "Speeding Up Neural Networks Execution: An Application to Speech Recognition", in Proc. of IEEE Workshop on Neural Networks for Signal Processing VI (NNSP-96), Kyoto, Japan 1996.
[10] R. Gemello, F. Mana, S. Scanzio, P. Laface, R. De Mori, *"Adaptation of Hybrid ANN/HMM models using hidden linear transformations and conservative training",* Proc. of Icassp 2006, Toulouse, France, May 2006
[11] L. Fissore, F. Ravera, P. Laface, "Acoustic-Phonetic Modeling for Flexible Vocabulary Speech Recognition", in Proc. of EUROSPEECH '95, Madrid, September 1995.
[12] Neto J. P., Martins C., Almeida L. B., "Speaker-Adaptation in a Hybrid HMM-MLP Recognizer", in *Proc. of ICASSP '96*.