Toward Noise–tolerant Acoustic Models

Edmondo Trentin¹, Marco Gori

Dip. di Ingegneria dell’Informazione
Università di Siena, V. Roma, 56 - Siena (Italy)
{trentin,marco}@dii.unisi.it

Abstract

State-of-the-art acoustic models, relying on hidden Markov models (HMMs), are heavily noise-sensitive: recognition performance drops whenever a significant difference in acoustic conditions holds between the training and the test environments. Practical, yet partial, attempts to tackle the problem are usually based on noise reduction via spectral subtraction, blind source separation, parameter adaptation/normalization, HMM retraining, etc. But the relevance of developing acoustic models that are intrinsically robust has to be stressed. Robustness to noise is related to the generalization capabilities of the model. Artificial neural networks (ANNs) appear to be a promising alternative, but they historically failed as a general paradigm for speech recognition [1]. This paper faces the problem by (i) investigating the recognition performance of the ANN/HMM hybrid proposed by the authors [2] over tasks with noisy signals, and (ii) proposing an explicit “soft” weight grouping technique, capable to improve its robustness. Experiments over noisy speaker-independent connected-digits strings are presented. In particular, results on the VODIS III/SpeechDatCar database, collected in a real car environment, show the dramatic gain over the standard HMM, as well as over Bourlard and Morgan’s hybrid.

1. Introduction

In recent years, noise-tolerance has become a requirement in several real-world automatic speech recognition (ASR) applications. In spite of that, robustness to noise and to variability in acoustic conditions are problems far from being solved in systems based on hidden Markov models (HMMs). We stress the fact that artificial neural networks (ANNs) learning theory draws a relationship between “learning with noise” and the generalization capabilities of the learning machine [3]. In [2] we introduced a novel hybrid acoustic model for ASR, based on the combination of ANNs and HMMs. The ANN/HMM hybrid is related to the paradigms proposed by Bourlard and Morgan [4] and by Bengio [5]. The former features an architecture where a Multilayer Perceptron (MLP) estimates probabilistic quantities (the so-called conditional transition probabilities) associated with individual states of the underlying HMM. Unfortunately, its training scheme is eventually heuristic. Bengio’s approach provides us with a formalism that forms the basis for the development of hybrid training algorithms aimed at the extremization of a global criterion function via joint optimization of the HMM parameters and of the ANN weights. Bengio’s architecture is based on a connectionist feature extractor for a standard HMM, i.e. the acoustic model is an HMM with its intrinsic limitations.

The novel hybrid relies on an HMM topology, including standard initial probabilities and transition probabilities \( a_{ij} \) for each pair of states \( i, j \), while the emission probabilities are estimated by an ANN. An output unit of the latter holds for each of the \( Q \) states in the HMM, with the understanding that \( i \)-th output value \( a_{oi}(t) \) at time \( t \) represents the emission probability \( b_{oi} \) for the corresponding \( (i\text{-th}) \) state, evaluated over current acoustic observation \( y_t \). While recognition is accomplished applying the usual Viterbi algorithm, novel training techniques are introduced. Learning rules for connection weights are calculated according to gradient-ascent to maximize a global criterion function, namely the likelihood \( P(Y \mid \lambda) \) of the observation sequence \( Y \) given the acoustic model \( \lambda \) under consideration (ML criterion). The ANN/HMM hybrid is expected to overcome major limitations of standard HMM-based ASR systems [4, 1], i.e., imposition of a specific parametric assumption for the emission probability density functions (pdfs), and limited generalization capabilities.

The scope of the paper is twofold: first, robustness, i.e., generalization properties of the novel ANN/HMM hybrid have to be evaluated in recognition tasks over noisy data. Then, an explicit technique aimed at improving generalization of the hybrid is developed and evaluated. In [6] we proposed an approach capable to improve learning and generalization capabilities in feed-forward ANNs. It relies on the introduction of a parameter \( \lambda \) within the training scheme, assuming activation functions in the form \( y = \lambda f(x) \). Supervised, gradient-descent training algorithms - having the goal to minimize the mean squared error over a labeled training set - were introduced to learn the \( \lambda \)'s from the data. In the present paper, activation functions in the form \( y = \lambda f(x) \) are allowed within the novel ANN/HMM hybrid paradigm.

We introduce gradient-ascent training algorithms to learn \( \lambda \) from the data according to the training criterion for the ANN/HMM hybrid, i.e. \( P(Y \mid \lambda) \). Individual units of the ANN may be subject to \( \lambda \) or not, and the latter may be layer-by-layer, or unit-specific. The algorithms are applied simultaneously with the ML learning rules for the connection weights [2]. The other parameters of the underlying HMM, namely initial and transition probabilities, are estimated via the Baum-Welch algorithm. A “soft” weight grouping [3] technique for the ANN/HMM is obtained, where all connection weights that start from a given unit (or set of units) having amplitude \( \lambda \) are grouped together. The range of their values is conditioned by \( \lambda \), and the latter is learned from the data according to contributions from all the weights in the group. As a consequence, different search paths within the weight space are induced. In [6, 2], a theoretical analysis is carried out to investigate the rationale behind the improvement yielded by the approach. It turns out that: (i) the grouping is motivated in terms of Bayesian evidence [3]; (ii) the technique can be described as a novel, gradient-driven

¹E.T. is also with ITC-irst - Pantè di Povo, Trento (Italy).
adaptive learning rate [3] scheme; (iii) the phenomenon of saturation of sigmoids [3] is avoided. Since the activations in the form $y = \lambda f(x)$ may be described as functions with an amplitude $\lambda$, a qualitative interpretation of the proposed algorithms can be given in terms of trainable amplitudes of activation functions in the ANN/HMM hybrid.

2. Maximum-Likelihood Weight-Grouping

Let us consider a feed-forward ANN with $L$ layers, namely $L_0$, $L_1$, $\ldots$, $L_L$, where $L_0$ is the input layer and $L_L$ is the output layer. We write $w_{j,i,t}$ to denote the connection weight between the $j$-th unit in layer $L_{t-1}$ and the $i$-th unit in layer $L_t$. Input to the latter unit at time $t$ is $a_{i,t}(t)$, and its output is $o_{i,t}(t)$, such that: $a_{i,t}(t) = \sum_{j \in L_{t-1}} w_{j,i,t} a_{j,t-1}(t)$. Sums over $k \in L_t$ are extended to all the indexes of units of layer $L_t$. An activation function $f_{i,t}()$ is associated with each unit, defining the input/output relationship $o_{i,t}(t) = f_{i,t}(a_{i,t}(t))$. The analytical form of the activation functions may be unit-specific. We assume that $f_{i,t}(\cdot)$ can be either in the form

$$f_{i,t}(a_{i,t}(t)) = \lambda_{i,t} f_i(o_{i,t}(t))$$

that is the case of functions that do have a learnable amplitude $\lambda_{i,t}$; or in the form

$$f_{i,t}(a_{i,t}(t)) = \tilde{f}_i(a_{i,t}(t))$$

that is the case of functions not depending upon any adaptive amplitudes. In the following the symbol $f_{i,t}(a_{i,t}(t))$ will refer to a function of $a_{i,t}(t)$ which does not explicitly depend on $\lambda_{i,t}$. The global criterion function $C$ to be maximized during training is the likelihood of the acoustic observations given the model:

$$C = \sum_{t \in T} o_{i,t}$$

where $T$ denotes the set of final states [5] for the HMM corresponding to the phonetic transcription of the observation sequence $Y = y_1, \ldots, y_T$ under consideration, and $o_{i,t}$ is the forward term at time $T$ for $i$-th state of the HMM. Three major cases may be taken into consideration.

Case 1: a single $\lambda$ is shared among all units of the ANN. Since all weights are grouped together, this is no longer a grouping technique and does not lead to significant improvement.

Case 2: a different value $\lambda$ of the amplitude is defined for each layer $L_t$. Derivation of the present algorithm is carried out in [2].

Case 3: different $\lambda_{i,t}$ are possibly defined for each unit $i$ in each layer $L_t$ of the ANN. Calculations for this case, that is the most general, are outlined in the following.

Hill climbing gradient-ascent prescribes a rule of the kind

$$\Delta \lambda_{i,t} = \eta \frac{\partial C}{\partial \lambda_{i,t}}$$

where $\eta$ is the learning rate, for each $t = 1, \ldots, L$ and for each $i \in L_t$ for which Eq. (1) holds. Observing that $\frac{\partial C}{\partial \lambda_{i,t}} = \frac{\partial C}{\partial o_{i,t}}$, and that $\frac{\partial o_{i,t}}{\partial \lambda_{i,t}} = \beta_{i,t}$ [5], we can write:

$$\frac{\partial C}{\partial \lambda_{i,t}} = \sum_{o \in o_{i,t}} \frac{\partial C}{\partial o} \frac{\partial o}{\partial \lambda_{i,t}} = \sum_{o \in o_{i,t}} \frac{\partial C}{\partial o} \lambda_{i,t} \beta_{i,t}$$

$$\frac{\partial C}{\partial o_{i,t}} = \sum_{o \in o_{i,t}} \frac{\partial C}{\partial o} \frac{\partial o}{\partial o_{i,t}} = \sum_{o \in o_{i,t}} \frac{\partial C}{\partial o} o_{i,t}$$

$$\frac{\partial o_{i,t}}{\partial \lambda_{i,t}} = \beta_{i,t}$$

$$\frac{\partial o_{i,t}}{\partial o_{i,t}} = 1$$

$$\frac{\partial o_{i,t}}{\partial o_{i,t}} = 0$$

First of all, let $t$ be an output unit, i.e., $t = L$. If $t \neq i$ then $\frac{\partial o_{i,t}(t)}{\partial \lambda_{i,t}} = 0$. Otherwise, we can write:

$$\frac{\partial o_{i,t}(t)}{\partial \lambda_{i,t}} = \tilde{f}_i(a_{i,t}(t))$$

Defining the quantity $\delta_{i,t}(i,t)$ as

$$\begin{cases} 1 & \text{if } t = L, i = t \\ 0 & \text{if } t = L, i \neq t \\ \sum_{j \in L_{t+1}} w_{j,i+1,t} \delta_{j,i+1,t} f_{j,i+1,t}(a_{j,i+1,t}(t)) & \text{otherwise} \end{cases}$$

we can rewrite Eq. (5) in the form:

$$\frac{\partial C}{\partial \lambda_{i,t}} = \tilde{f}_i(a_{i,t}(t)) \sum_{o \in o_{i,t}} \frac{\partial C}{\partial o} o_{i,t} \lambda_{i,t} \beta_{i,t}$$

Let us now consider the case of units in the first hidden layer, that is to say $t = L - 1$. Taking partial derivative of $o_{i,t}(t)$ w.r.t. $\lambda_{t-1}$ and repeatedly applying the chain rule yield:

$$\frac{\partial C}{\partial \lambda_{i,t}} = \frac{\partial C}{\partial o_{i,t}(t)} \frac{\partial o_{i,t}(t)}{\partial \lambda_{i,t}}$$

Applying again the chain rule:

$$\frac{\partial C}{\partial o_{i,t}(t)} = \frac{\partial C}{\partial o_{i,t}(t)} \frac{\partial o_{i,t}(t)}{\partial \lambda_{i,t}}$$

Substituting Eq. (8) into Eq. (7) we have:

$$\frac{\partial C}{\partial \lambda_{i,t}} = \sum_{o \in o_{i,t}} \frac{\partial C}{\partial o} o_{i,t} \lambda_{i,t} \beta_{i,t}$$

It is possible to prove by induction [2] that a similar formulation holds also in the case of units in the subsequent layers, i.e., $t < L - 1$. In summary, Eqs. (6) and (10) can be substituted into Eq. (4), obtaining an on-line learning rule in the common form:

$$\Delta \lambda_{i,t} = \eta \sum_{t = 1}^{L} \sum_{i = 1}^{L} \beta_{i,t} \frac{\partial o_{i,t}(t)}{\partial \lambda_{i,t}}$$

for each layer $L_t = L_1, \ldots, L_L$ in the ANN, and for each unit $i$ in $L_t$.\(^1\)

\(^1\)Taking the partial derivatives of the left- and right-hand sides of the definition of forward terms, i.e. $o_{i,t} = \sum_{j \in L_{t-1}} a_{j,t-1} w_{j,i,t}$, w.r.t. $b_{i,t}$.\(^2\)
3. Experiments

Two experimental setups are considered. Both of them rely on the same training step, accomplished on the training part of cdigits from the SPK database. SPK is a clean corpus, i.e., it was collected at ITC-irst (Trento, Italy) under laboratory conditions, using a close-mouth microphone. This training set includes 500 utterances of Italian connected 8-digit strings (4000 words in total) from 20 speakers (10 male and 10 female), sampled at a rate of 16kHz. The feature space was defined by using 20ms Hamming windows, with an overlap of 10ms, and extracting 8 Mel Frequency Scaled Cepstral Coefficients (MFSCCs) and the log-energy of the signal.

The difference between the two experimental setups concerns the test sets, where no further training, re-training or adaptation of model parameters is accomplished: our goal is the evaluation and comparison of the robustness of the acoustic models themselves. In the first instance (Section 3.1), noise is added to the clean signals of the test part of cdigits. This allows for an evaluation of the acoustic models at different Signal to Noise Ratios (SNRs), preserving the same conditions (e.g., the transducer) that underly the acquisition of the speech signals in the training set. In the second instance (Section 3.2), the models trained over the same clean training set are applied to the recognition task of connected digits strings acquired in a real car environment.

The HMM and ANN/HMM topologies feature 11 left-to-right word models (one per digit, plus one for the “silence” model), with a number of states equal to the length of the phonetic transcription of each Italian digit according to the SAMPA (Speech Assessment Methods Phonetic Alphabet)3 acoustic-phonetic units. The HMM contains 8 Gaussian pdfs per state, and the Segmental k-Means initialization, Baum-Welch training and Viterbi decoding algorithms are used. The ANN was chosen according to the experimental guidelines pointed out in [2]: it is a 2-layer MLP with a 93-sigroids hidden layer and a 40-sigroids output layer, initialized according to the Bourlard and Morgan-like iterative BP/Viterbi scheme [4]. Since the number of free parameters in the different models belongs to the same magnitude, the comparison in terms of generalization (bias vs. variance) [3] is fair, from a learning theory point of view.

3.1. Experiments with the additive-noise model

The portion of the SPK corpus used herein is the test part of cdigits. It consists of other 500 clean utterances (4000 words in total) from 20 speakers (11 male and 9 female) not involved in the training process. Real noise collected in an office environment was added to the clean speech signal, assuming an additive model for the overall signal $s(t)$ at time $t$ in the form $s(t) = x(t) + n(t)$, where $x(t)$ is the clean signal and $n(t)$ is the additive noise. The technique described in [7] was applied in order to compute the additive noise simulation.

Table 1 shows the results obtained with the standard HMM, Bourlard and Morgan’s hybrid, and the present ANN/HMM hybrid (with fixed, as well as adaptive $\lambda$’s), over the clean test set and over the noisy test set (at a reference SNR of 20dB). Results are expressed in terms of String Recognition Rate (SRR) and Word Recognition Rate (WRR). The SRR is obtained by evaluating the error rate at the word-strings level (i.e., the fraction of test sequences that are recognized without any errors).

The performance of the standard HMM drops when the recognizer is fed with noisy signals. Bourlard and Morgan’s hybrid, that basically performs as the HMM over the clean data, turns out to be more robust to noise, due to the generalization capabilities of the MLP. In the present experiments, Bourlard and Morgan’s paradigm uses the same ANN topology as that of the novel ANN/HMM hybrid. The latter (with fixed amplitudes) provides us with a significant gain in terms of WRR, proving itself much more noise-tolerant than the standard models.

Let us analyze the results obtained from the introduction of trainable $\lambda$’s. Results on the clean speech could suggest that the different variants of the hybrid are all very close to the global maximum given the present architecture (or, at least, that they converged to highly similar solutions nearby a common local maximum in the parameter space). Recognition performance on the noisy speech signals points out that this is not the case: Cases 2 and 3 induce weight groupings that allow for the realization of more robust acoustic models. This fact shows that different maxima of the criterion function are reached, i.e. different solutions in the parameter space that allow for improved generalization. In particular, Case 3 yields a 64.58% relative Word Error Rate (WER) reduction with respect to the standard HMM, and a 17.58% relative WER reduction w.r.t. the ANN/HMM with fixed $\lambda$. Fig. 1 plots the WRR curves for the HMM, the ANN/HMM and the hybrid with unit-by-unit adaptive amplitudes, as the amount of noise in the signal increases.

![Figure 1: WRR curves as functions of the SNR for: standard HMM (lower curve), proposed ANN/HMM hybrid (middle curve), and ANN/HMM with Case 3 of adaptive amplitudes (upper curve).](image-url)

It is seen that: (i) the HMM rapidly worsens its performance in the presence of noise; at a SNR of 40dB (rather light noise) it scores a 87.17% WRR, i.e. about 3% absolute WRR loss w.r.t. its performance over the clean signals; (ii) the ANN/HMM hybrid shows a more stable and noise-tolerant behavior: at a SNR of 30dB it is still well over the performance of the HMM evaluated on the clean signals, at 20dB the absolute gap in WRR between the two models is 17.32 points; (iii) the ANN/HMM with adaptive $\lambda$’s is constantly higher than the bare ANN/HMM, and the difference becomes significant when the noise level grows over a SNR of 30dB. At the latter SNR level, its performance is comparable with that yielded by the bare hybrid evaluated on the clean data, while up to 20dB it compares favorably with the performance of the standard HMM on the clean signals.

---

3SPK is available from the European Language Resources Association (ELRA).

3SAMPA is a machine-readable phonetic alphabet, originally developed under the ESPRIT project SAM (Speech Assessment Methods).

---
Table 1: String Recognition Rate (SRR) and Word Recognition Rate (WRR) over the clean and noisy (SNR: 20dB) test signals of SPK.

<table>
<thead>
<tr>
<th>Acoustic Model</th>
<th>Clean Signals SRR (%)</th>
<th>Noisy Signals SRR (%)</th>
<th>Clean Signals WRR (%)</th>
<th>Noisy Signals WRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard HMM</td>
<td>46.60</td>
<td>10.20</td>
<td>47.40</td>
<td>24.40</td>
</tr>
<tr>
<td>Bourlard and Morgan’s model</td>
<td>68.60</td>
<td>41.40</td>
<td>68.20</td>
<td>44.00</td>
</tr>
<tr>
<td>Novel ANN/HMM hybrid, fixed amplitudes</td>
<td>68.60</td>
<td>41.40</td>
<td>68.20</td>
<td>44.00</td>
</tr>
<tr>
<td>ANN/HMM hybrid, layer-by-layer adaptive (\lambda_i)'s</td>
<td>68.40</td>
<td>47.00</td>
<td>86.80</td>
<td>88.33</td>
</tr>
</tbody>
</table>

3.2. ASR in a real car environment

The present noisy corpus, collected under the VODIS-II/SpeechDatCar projects[8] in a real car environment, is available from ELRA. It consists of 200 speakers (100 male and 100 female) recorded under a variety of noisy car conditions (e.g., traffic, different car speed, various road/highway and weather conditions, etc.). Each speaker uttered variable-length, continuous Italian digit-strings, during two distinct sessions, for a total of 2700 utterances. Each string has variable length, ranging from a minimum of 1 to a maximum of 16 words, with an average length of about 6 words.

Signals were acquired using a close-mouth microphone placed in front of the driver. It is worth underlining that, while the 20 speakers from the SPK database used for training are mostly from the North-East of Italy, and are aged in the range 25-40, the VODIS-II/SpeechDatCar corpus was designed to cover the Country in a rather uniform manner (i.e., a variety of regional accents is represented), along with a much wider range coverage of the age of speakers (from very young drivers to old man/women). These facts deepen the gap between the training and test conditions, that are no longer limited to the presence of noise as in the previous section. Further differences are brought in by the use of a different microphone, and by the spontaneous changes in vocal tracts induced in the speaker by the acoustic conditions of the environment, mainly due to the Lombard effect.

The first 8 MFSCCs and the signal log-energy were extracted as usual, in order to allow application of the same acoustic models trained above. Table 2 reports the results. The same conclusions drawn in the previous section still hold, as far as concerns the comparison among the different models, their noise-tolerance, etc.

Table 2: SRR and WRR over the noisy test signals of the VODIS II/SpeechDatCar corpus.

<table>
<thead>
<tr>
<th>Acoustic Model</th>
<th>SRR (%)</th>
<th>WRR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard HMM</td>
<td>37.58</td>
<td>71.93</td>
</tr>
<tr>
<td>Bourlard and Morgan’s model</td>
<td>41.80</td>
<td>73.96</td>
</tr>
<tr>
<td>ANN/HMM, fixed amplitudes</td>
<td>53.68</td>
<td>80.58</td>
</tr>
<tr>
<td>ANN/HMM, layer-by-layer adaptive (\lambda_i)'s</td>
<td>56.50</td>
<td>82.74</td>
</tr>
<tr>
<td>ANN/HMM, unit-by-unit (\lambda_i)'s</td>
<td>57.02</td>
<td>83.07</td>
</tr>
</tbody>
</table>

4 Part of SpeechDatCar and Vodis-II projects is funded by the Commission of the EC, Telematics Applications Programme, Language Engineering, Contracts LE4-8334 and LE4-8336.

4. Conclusion

One of the most severe limitations in standard HMMs is their intrinsically limited robustness to noise. In addition, they actually lack of ad hoc training schemes (e.g. regularization techniques) capable to tackle such problem by explicitly improving their generalization capabilities. In fact, this problem has a strong negative influence on the test recognition performance under (noisy) acoustic conditions that differ from those that characterize the training set. A "soft" weight-grouping technique for a novel ANN/HMM hybrid was proposed as an effective direction toward the solution of these problems, i.e. an acoustic model that turns out to be much more noise-tolerant than standard HMMs in the experiments presented herein.

5. References