

Evaluation on the Aurora 2 Database of Acoustic Models that are less Noise-sensitive

Edmondo Trentin

Dip. di Ingegneria dell'Informazione
Univ. di Siena, V. Roma, 56 - Siena (Italy)
trentin@dii.unisi.it

Marco Matassoni

ITC-irst
Via Sommarive, 18 - Povo (Italy)
matasso@itc.it

Marco Gori

Dip. di Ingegneria dell'Informazione
Univ. di Siena, V. Roma, 56 - Siena (Italy)
marco@dii.unisi.it

Abstract

The Aurora 2 database may be used as a benchmark for evaluation of algorithms under noisy conditions. In particular, the *clean training/noisy test* mode is aimed at evaluating models that are trained on clean data only without further adjustments on the noisy data, i.e. under severe mismatch between the training and test conditions. While several researchers proposed techniques at the front-end level to improve recognition performance over the reference hidden Markov model (HMM) baseline, investigations at the back-end level are sought. In this respect, the goal is to develop acoustic models that are intrinsically less noise sensitive. This paper presents the word accuracy yielded by a non-parametric HMM with connectionist estimates of the emission probabilities, i.e. a neural network is applied instead of the usual parametric (Gaussian mixture) probability densities. A regularization technique, relying on a maximum-likelihood parameter grouping algorithm, is explicitly introduced to increase the generalization capability of the model and, in turn, its noise-robustness. Results show that a 15, 43% relative word error rate reduction w.r.t. the Gaussian-mixture HMM is obtained by averaging over the different noises and SNRs of Aurora 2 test set A.

1. Introduction

Designed as a common platform to evaluate the DSR (distributed speech recognition) approach, as well as for developing a coding standard for acoustic parameters, the Aurora 2 database was proposed in [1] for the comparison of speech recognition algorithms and signal-processing techniques aimed at improving performance under noisy conditions. Although new releases of the Aurora database are now available [2], Aurora 2 still remains a popular and severe evaluation benchmark. As pointed out in [1], the different recognition tasks involved in Aurora 2 are intended as a challenge for both the acoustic front-end (parameter extraction, noise filtering, etc.) and the back-end (the acoustic model), commonly based on hidden Markov models (HMMs). Indeed, several researchers proposed techniques at the front-end level to improve recognition performance over the reference HMM baseline, e.g. [2]. Less emphasis has been placed on the back-end: can we strengthen the acoustic model in order to reduce its noise sensitivity whenever training and test conditions are different? One of the Aurora 2 subtasks is particularly suitable to address the question, namely the *clean training/noisy test* mode (briefly reviewed in Section 3.1). The latter is aimed at evaluating models that are trained on clean data only and tested on different noise sources at various signal-to-noise ratios (SNRs), without any further adjustments of model parameters.

This paper presents the word accuracy yielded by a non-

parametric HMM with connectionist estimates of the emission probabilities, i.e. a neural network is applied instead of the usual parametric (Gaussian mixture) probability densities. Results are compared w.r.t. a Gaussian-mixture HMM on the *test set A* part of Aurora 2. In [3] we introduced a novel hybrid acoustic model, based on the combination of artificial neural networks (ANNs) and HMMs. A simple architecture is shown in Figure 1. This ANN/HMM 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 states in the HMM, with the understanding that i -th output value $o_i(t)$ at time t represents the emission probability $b_{i,t}$ for the corresponding (i -th) state, evaluated over current acoustic observation \mathbf{y}_t . Recognition is accomplished applying the usual Viterbi algorithm, while a novel maximum-likelihood (ML) global training technique was introduced. The algorithm, reviewed in Section 2, relies on gradient-ascent to maximize the likelihood $L = P(Y | \mathcal{M})$ of the acoustic observation sequence Y given the model \mathcal{M} under consideration. Differences w.r.t. previous ANN/HMM hybrids were pointed out in [4, 3].

The ANN/HMM is expected to overcome major limitations of standard HMMs [5], i.e., imposition of a specific parametric assumption for the emission probability density functions (*pdfs*), and limited generalization capabilities. As a matter of fact, ANNs learning theory draws a relationship between "learning with noise" and the *generalization* capabilities of the learning machine [6]. Application of robust training techniques (e.g. regularization) during the learning process on clean (non-noisy) data is aimed at improving generalization, reducing overfitting, resulting in improved performance on noisy test data.

A regularization technique, based on a ML parameter grouping algorithm, is thus explicitly given to increase the generalization capability of the model and, in turn, its noise-robustness. In [7] we proposed such an approach relying on the introduction of a parameter λ within the training scheme, assuming activation functions in the form $y = \lambda f(x)$. The gradient-ascent training algorithm to learn λ from the data according to the training criterion for the ANN/HMM hybrid, i.e. the likelihood L , is reviewed in Section 2. A "soft" parameter grouping [6] technique for the ANN/HMM is obtained, where all connection weights that start from a given unit (or set of units) subject to λ are *grouped* together. The *range* of their values is conditioned by λ , 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 [8], a theoretical analysis is carried out to investigate the rationale behind the improvement yielded by the approach. The algorithm is applied simultaneously with the ML learning rule

for the connection weights. The other parameters of the underlying HMM, namely initial and transition probabilities, are estimated via the Baum-Welch algorithm. Experimental results (Section 3.3) show that a 15, 43% relative word error rate reduction w.r.t. the Gaussian-mixture HMM is obtained by averaging over the different noises and SNRs of Aurora 2 test set A.

2. Review of the proposed acoustic model

In this Section, a sketch of the fundamental calculations for ML training of the ANN/HMM are given. Along the line of standard Gaussian-density HMM parameter-estimation algorithms, the global criterion function C to be maximized by the model during training is the *likelihood* L of the acoustic observations given the model: $C = L$, where:

$$L = \sum_{i \in \mathcal{F}} \alpha_{i,T} \quad (1)$$

and the α 's are the usual *forward* probabilities. The sum is extended to the set \mathcal{F} of all possible *final* states within the HMM [9] corresponding to the current phonetic transcription, which is supposed to involve Q states, and T is the length of the current observation sequence $Y = \mathbf{y}_1, \dots, \mathbf{y}_T$.

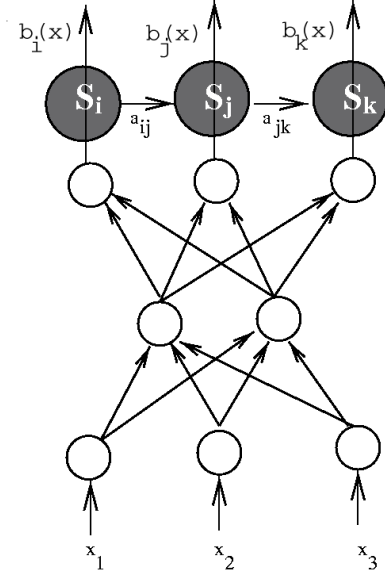


Figure 1: Example of a simple ANN/HMM architecture with connectionist estimation of emission probabilities for the underlying HMM.

Assuming that the ANN represents the emission *pdfs* in a proper manner, and given a generic weight w of the ANN, hill-climbing gradient-ascent over C prescribes a learning rule of the kind:

$$\Delta w = \eta \frac{\partial L}{\partial w} \quad (2)$$

where η is the *learning rate*. Let us observe (after [9]) that the following property can be easily shown to hold true:

$$\frac{\partial \alpha_{i,t}}{\partial b_{i,t}} = \frac{\alpha_{i,t}}{b_{i,t}}. \quad (3)$$

According to [10, 9], the following theorem can be proved to hold true: $\frac{\partial L}{\partial \alpha_{i,t}} = \beta_{i,t}$, for each $i = 1, \dots, Q$ and for each

$t = 1, \dots, T$. Given the theorem and Equation (3), repeatedly applying the chain rule we can expand $\frac{\partial L}{\partial w}$ by writing:

$$\begin{aligned} \frac{\partial L}{\partial w} &= \sum_i \sum_t \frac{\partial L}{\partial b_{i,t}} \frac{\partial b_{i,t}}{\partial w} \\ &= \sum_i \sum_t \beta_{i,t} \frac{\alpha_{i,t}}{b_{i,t}} \frac{\partial b_{i,t}}{\partial w}. \end{aligned} \quad (4)$$

where the sums are extended over all states $i = 1, \dots, Q$ of the HMM corresponding to the correct transcription of the training utterance under consideration, and to all $t = 1, \dots, T$, respectively. Let us consider a multilayer Perceptron (MLP), the j -th output of which, computed over t -th input observation \mathbf{y}_t , is interpreted as a non-parametric estimate of the emission probability $b_{j,t}$ associated with j -th state of the HMM at time t . An activation function $f_j(x_j(t))$ is associated with each unit j of the MLP, where $x_j(t)$ denotes input to the unit itself at time t . The corresponding output $o_j(t)$ is given by $o_j(t) = f_j(x_j(t))$. This ANN is assumed to have L layers $\mathcal{L}_0, \mathcal{L}_1, \dots, \mathcal{L}_L$, where \mathcal{L}_0 is the input layer, and \mathcal{L}_L is the output layer. For notational convenience we write $i \in \mathcal{L}_k$ to denote the index of i -th unit in layer \mathcal{L}_k .

Given a generic weight w_{jk} between k -th unit in layer \mathcal{L}_{l-1} and j -th unit in layer \mathcal{L}_l , and defining the quantity

$$\delta_j(i, t) = \begin{cases} f'_j(x_j(t)) & \text{if } l = L, i = j \\ 0 & \text{if } l = L, i \neq j \\ f'_j(x_j(t)) \sum_{n \in \mathcal{L}_{l+1}} w_{ni} \delta_n(j, t) & \text{otherwise} \end{cases} \quad (5)$$

for each $i \in \mathcal{L}_n$, it is possible to show that [3]:

$$\frac{\partial b_{i,t}}{\partial w_{jk}} = \delta_j(i, t) o_k(t). \quad (6)$$

Using the above calculations to expand Equation (4) and substituting it into Equation (2), the latter can now be restated in the form of a *learning rule* for weight w_{jk} , by writing:

$$\Delta w_{jk} = \eta \sum_{i=1}^Q \sum_{t=1}^T \beta_{i,t} \frac{\alpha_{i,t}}{b_{i,t}} \delta_j(i, t) o_k(t) \quad (7)$$

where the term $\delta_j(i, t)$ is computed via Equation (5).

The parameter-grouping scheme is obtained by introducing *trainable* parameters $\lambda_{i,\ell}$ for each unit i in each layer \mathcal{L}_ℓ , and by considering activation functions in the form

$$f_{i,\ell}(x_{i,\ell}(t)) = \lambda_{i,\ell} \tilde{f}_{i,\ell}(x_{i,\ell}(t)) \quad (8)$$

where dependence of the different quantities on the specific layer \mathcal{L}_ℓ was explicitly stated for notational convenience in the calculations. In the following, the symbol $\tilde{f}_{i,\ell}(x_{i,\ell}(t))$ will refer to a function of $x_{i,\ell}(t)$ which does not explicitly depend on $\lambda_{i,\ell}$.

Considering criterion (1), and relying on gradient ascent as in Eq. (4), we have:

$$\frac{\partial C}{\partial \lambda_{i,\ell}} = \sum_i \sum_t \beta_{i,t} \frac{\alpha_{i,t}}{b_{i,t}} \frac{\partial b_{i,t}}{\partial \lambda_{i,\ell}}. \quad (9)$$

Again, being $\frac{\partial b_{i,t}}{\partial \lambda_{i,\ell}} = \frac{\partial o_{i,L}(t)}{\partial \lambda_{i,\ell}}$, by defining the quantity $\delta_{i,\ell}(i, t)$ as

$$\begin{cases} 1 & \text{if } \ell = L, \ell = i \\ 0 & \text{if } \ell = L, \ell \neq i \\ \sum_{j \in \mathcal{L}_{\ell+1}} w_{j,\ell,\ell+1} \delta_{j,\ell+1}(i, t) f'_{j,\ell+1}(x_{j,\ell+1}(t)) & \text{otherwise} \end{cases} \quad (10)$$

it is possible to prove by induction [4] that:

$$\frac{\partial \mathcal{O}_{i,\ell}(t)}{\partial \lambda_{i,\ell}} = \delta_{i,\ell}(i, t) \tilde{f}_{i,\ell}(x_{i,\ell}(t)). \quad (11)$$

In summary, Eq. (11) can be substituted into Eq. (9), obtaining an on-line learning rule in the form:

$$\Delta \lambda_{i,\ell} = \eta \sum_{i=1}^Q \sum_{t=1}^T \beta_{i,t} \frac{\alpha_{i,t}}{b_{i,t}} \delta_{i,\ell}(i, t) \tilde{f}_{i,\ell}(x_{i,\ell}(t)) \quad (12)$$

for each layer $\mathcal{L}_\ell = \mathcal{L}_1, \dots, \mathcal{L}_L$ in the ANN, and for each unit i in \mathcal{L}_ℓ .

3. Experiments

3.1. The dataset

As reported in [1], the Aurora 2 relies on the training part of the TIDigits database. The *clean* training data include 8440 utterances of English isolated digits and connected (up to 7) digit strings from US-American adults (55 male and 55 female), downsampled to a rate of 8kHz. The feature space used herein was defined by taking 20ms Hamming windows, with an overlap of 10ms, and extracting 8 *Mel Frequency Scaled Cepstral Coefficients* (MFSCCs) and the log-energy of the signal. First-order and second-order time derivatives of the static features were calculated, as well, resulting in a 27-dim feature space.

According to [1], the Aurora 2 noisy *test set A* is defined by splitting 4004 utterances (52 male and 52 female) from the test part of the TIDigits database into 4 subsets, each having 1001 utterances. 4 noise sources (suburban train, crowd of people or babble, car, exhibition hall) were added to each subset at different SNRs. In the present experiments we consider SNRs ranging from 0 to 20 dB, i.e. we deal with 5 instances of each of the 4 subsets, for a total of 20020 test utterances.

No further training, re-training or adaptation of model parameters is accomplished on the noisy data: our goal is the evaluation and comparison of the robustness of the acoustic models themselves.

3.2. Topology of the models

The HMM and ANN/HMM topologies feature 12 left-to-right word models (one per digit - including the double pronunciation for 0 - plus one HMM for the “pause” model), with a number of states proportional to the length of a standard phonetic transcription of each English digit, and a single state for the pause model (78 states total). No skip-state transitions were introduced. The HMM contains 8 Gaussian *pdfs* per state with diagonal covariance matrices, and the Segmental k-Means initialization, Baum-Welch training and Viterbi decoding algorithms are used.

The ANN was chosen accordingly: it is a 2-layer MLP with a 180-sigmoids hidden layer and a 78-sigmoids output layer, i.e. one output unit for each one of the states in the underlying HMM. The ANN was initialized according to the Bourlard and Morgan-like iterative BP/Viterbi scheme [11, 4]. Since the

number of free parameters in the different models belongs to the same *magnitudo*, the comparison in terms of generalization (bias vs. variance) [6] is fair, from a learning theory point of view.

3.3. Results

Table 1 shows the word accuracy obtained with the Gaussian-density HMM developed by the Speech Group at ITC-irst [12]. The results are not perfectly comparable with those reported in [1, 2]; this is likely to be due to: (a) the different topologies (longer left-to-right HMMs are used therein); (b) the different number of HMMs (two models are used in [1] to model the “silence” at the end/tail of utterances, and the “pause” between pairs of words); (c) the absence of any pre-segmentation of signals (e.g., start-end point detection) herein; and (d) the lower dimensionality of the feature space (8 MFSCCs instead of 12).

Table 1: *Word accuracy (%) for Aurora 2 Test Set A in clean training mode using the present Gaussian-density HMM back-end.*

SNR	Subway	Babble	Car	Exhibtn.	Avg.
20 dB	93.37	93.62	94.15	91.61	93.19
15 dB	85.39	83.10	80.73	83.52	83.19
10 dB	58.55	59.76	48.37	53.75	55.11
5 dB	32.36	37.39	24.46	23.45	29.42
0 dB	14.92	16.99	9.10	10.71	12.93
Avg.	56.92	58.17	51.36	52.61	54.77

The results obtained with the ANN/HMM (without parameter grouping) are shown in Table 2, while Table 3 reports on the experiments with the ANN/HMM with parameter grouping. A graphical representation of the behavior of the back-ends in terms of average word accuracy¹ as a function of increasing noise is given in Figure 2.

Table 2: *Word accuracy (%) for Aurora 2 Test Set A in clean training mode using the ANN/HMM back-end without parameter grouping.*

SNR	Subway	Babble	Car	Exhibtn.	Avg.
20 dB	93.51	94.12	94.42	92.72	93.70
15 dB	90.03	88.25	87.36	89.71	88.84
10 dB	66.17	65.10	54.69	59.99	61.49
5 dB	38.45	43.02	30.22	29.58	35.32
0 dB	19.06	21.34	14.61	14.37	17.35
Avg.	61.44	62.37	56.26	57.27	59.34

It is seen that: (i) the ANN/HMM back-ends are less noise-sensitive than the Gaussian-mixture HMM, yielding improvement in terms of word accuracy in a systematic manner; (ii) the parameter grouping technique actually improves generalization capabilities of the model, i.e. its noise-robustness, improving performance over all test sets with an exception only, namely the “car” noise at a SNR of 20dB. In particular, by comparing the recognition performance averaged over all noise types and SNRs, a 15, 43% relative word error rate reduction w.r.t. the

¹ Averaged over the four different noise types at a given SNR.

Table 3: Word accuracy (%) for Aurora 2 Test Set A in clean training mode using the ANN/HMM back-end with parameter grouping.

SNR	Subway	Babble	Car	Exhibitn.	Avg.
20 dB	94.01	94.76	94.36	93.80	94.23
15 dB	91.14	90.75	89.24	91.08	90.55
10 dB	70.49	69.39	59.83	62.40	65.53
5 dB	41.06	45.58	34.98	35.17	39.20
0 dB	21.41	22.66	15.90	17.04	19.25
Avg.	63.62	64.63	58.86	59.90	61.75

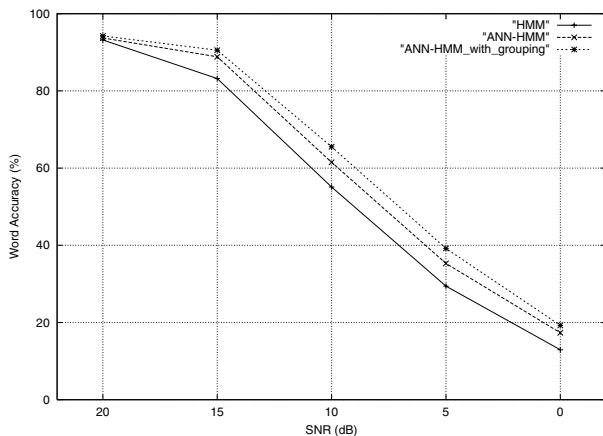


Figure 2: Average word accuracy (%) curves as functions of the SNR (dB) for the different back-ends: HMM with Gaussian mixtures (lower curve), proposed ANN/HMM (middle curve), and ANN/HMM with parameter grouping (upper curve).

Gaussian-based HMM is obtained. In addition, the ANN/HMM with parameter grouping compares favorably with the baseline results presented in [1, 2], too, in spite of the fact that slightly more complex HMM topologies were applied therein. This represents a better starting point than that provided by the Gaussian-based HMM toward the development of less noise-sensitive recognition systems via the combination with robust acoustic front-ends. The average word accuracy yielded by the different acoustic models are summarized in Table 4.

Table 4: Summary of average word accuracy yielded by the different acoustic back-ends, averaged over all noise conditions and SNRs.

Acoustic model	Avg. word accuracy
HMM with Gaussian mixtures	54.77 %
ANN/HMM	59.34 %
ANN/HMM with grouping	61.75 %

4. Conclusion

The Aurora 2 clean training/noisy test mode poses a challenge for acoustic back-ends: can we strengthen the acoustic model to reduce its noise sensitivity whenever training and test condi-

tions are different? While state-of-the-art approaches to speech recognition under noisy conditions stress robust front-ends to be applied along with standard HMMs, back-ends that are less noise-sensitive are hardly exploited. Indeed, one of the most severe limitations in Gaussian-density HMMs is their intrinsically limited robustness to noise. Furthermore, they lack of regularization schemes capable to tackle such drawback by improving their generalization capabilities. These problems have a negative influence on the test recognition performance under (noisy) acoustic conditions that differ from those that characterize the training set. A ML parameter-grouping technique for a novel ANN/HMM hybrid was evaluated on the Aurora 2 Test Set A. Experiments show that it turns out to be a promising alternative toward the solution of these problems. Current research is focused on combining the back-end with proper *ad-hoc* connectionist front ends, realizing an overall recognition system that is more robust to noise.

5. References

- [1] H.G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA*, 2000.
- [2] D. Macho, L. Mauuary, B. Noe, Y.M. Cheng, D. Ealey, D. Jovet, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front end on Aurora databases," in *Proc. ICSLP*, 2002.
- [3] E. Trentin and M. Gori, "Continuous speech recognition with a robust connectionist/markovian hybrid model," in *Proceedings of ICANN*, Vienna, Austria, August 2001.
- [4] E. Trentin, *Robust Combination of Neural Networks and Hidden Markov Models for Speech Recognition*, Ph.D. thesis, DSI, Univ. di Firenze, 2001.
- [5] E. Trentin and M. Gori, "A survey of hybrid ANN/HMM models for automatic speech recognition," *Neurocomputing*, vol. 37, no. 1-4, pp. 91-126, March 2001.
- [6] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, 1995.
- [7] E. Trentin and M. Gori, "Toward noise-tolerant acoustic models," in *Proceedings of Eurospeech 2001*, Aalborg, Scandinavia, September 2001.
- [8] E. Trentin, "Networks with trainable amplitude of activation functions," *Neural Networks*, vol. 14, no. 4-5, pp. 471-493, May 2001.
- [9] Y. Bengio, *Neural Networks for Speech and Sequence Recognition*, International Thomson Computer Press, London, UK, 1996.
- [10] J.S. Bridle, "Alphanets: a recurrent 'neural' network architecture with a hidden Markov model interpretation," *Speech Communication*, vol. 9, no. 1, pp. 83-92, 1990.
- [11] H. Bourlard and N. Morgan, *Connectionist Speech Recognition. A Hybrid Approach*, vol. 247, Kluwer Academic Publishers, Boston, 1994.
- [12] B. Angelini, F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, and M. Omologo, "Speaker independent continuous speech recognition using an acoustic-phonetic italian corpus," in *Proceedings of ICSLP*, 1994, vol. 3, pp. 1391-1394.