ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

Introduction to Speech Recognition Using Neural Networks *

Chris J. Wellekens

Communications Multimedia
Institut Eurécom
F-06904 Sophia-Antipolis

Abstract. As an introduction to a session dedicated to neural networks in speech processing, this paper describes the basic problems faced with in automatic speech recognition (ASR). Representation of speech, classification problems, speech unit models, training procedures and criteria are discussed. Why and how neural networks lead to challenging results in ASR is explained.

1. Introduction

This paper is an introduction to a special session dedicated to Speech Processing using neural networks and provides the basic knowledge in automatic speech recognition (ASR) to cognitive scientists with the aim to bridge the gap between communities. Only speech recognition is covered in this tutorial although the classification and mapping properties of MLP are used for many other applications in speech processing f.i. keyword spotting, speech synthesis, enhancement and noise robustness, speaker adaptation, speaker recognition, voiced/unvoiced/silence detection....

For more than ten years [1-4], neural networks have been used in ASR with scores comparable with those reached by traditional recognizers but with a simpler architecture and less parameters. Some neural network approaches are now challenging [5-6,26].

Successively, the paper shows how speech is represented for recognition tasks, then the principle of comparison of distorted sequences is explained on template models. The importance of hidden Markov models (HMM) is stressed and details are given on the training criteria. Eventually, the role of MLP and TDNN in speech recognition is explained and hybrid models show how to fully take advantage of the time alignment capabilities of HMM and of the discriminating properties of MLP.

*Eurecom's research is partially supported by its industrial partners: Ascom, Cegetel, France Telecom, Hitachi, IBM France, Motorola, Swisscom, Texas Instruments and Thomson CSF

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

2. Speech representation

Speech signal is produced by air flowing through the vocal tract articulated under brain control. After anti-aliasing filtering, the microphone signal is sampled at a frequency between 8 kHz (phone applications) and 16 kHz. The vocal tract is thus a time-varying system producing a non-stationary signal. As a consequence, speech signal is analyzed on short-time windows the duration of which is defined by the time constants of the articulatory apparatus (order of 10 ms). Contrary to the analysis of images, speech analysis is essentially based on harmonic analysis. There are indeed very few informations that can be immediately used for recognition in the waveform: this is mainly due to the mixture of what depends on the speech content together with what depends on the speaker or on the prosody (speed and intonation). Several analyses in the frequency domain that allow to some extent separation of speech content from irrelevant informations, are described [7-8] a.o. : filter bank analysis, smoothed spectrum, cepstral analysis. Linear prediction coding (LPC) provides a production model and is closely related to the harmonic analysis. Relying to the impressive capabilities of the human ear, different improvements have been brought to the front end. The first one is the use of a logarithmic scale for the frequency called Mel scale [8-9]. Speech representation plays a prominent role in the robustness to noise and to channel distortion. Filter banks [9] have non uniform bandwidths but critical bands related to the ear sensitivity and to the masking effect due to extra signals or noise (reduction ot ear sensitivity at a frequency due to the presence of signal power in the same critical band). PLP and RASTA-PLP [9] analyses have also been introduced to improve the behavior of speech recognition over the phone. They are all related to physiological observations on speech hearing.

In conclusion, analysis results in a sequence of vectors

$$X \stackrel{\text{def}}{=} \{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N\}$$

containing speech features at a period of 10ms.

Very soon, it has been observed that the dynamics of these vectors must be taken into account to improve the recognition scores. Feature vectors are thus extended by the differences between adjacent ones (Δ-features) and "acceleration" is also used ($\Delta\Delta$-features)[10].

3. Template-based recognition

To recognize a word, two main approches can be considered.

The first one is AI-oriented: experts build knowledge sources by analyzing characteristic features of words and these characteristics are looked for in the feature vector sequences of test words. Very soon, it was noticed that the high variability of speech waveforms made this approach unrealistic for large vocabularies or speaker independent recognizers.

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

The other approach is based on comparison between a prerecorded sequence of vectors representing a word utterance (template) and the test utterance. Indeed, there is usually a strong distortion of the time scale between test and reference utterances. Not only the speaking rate may be different but also the length variation of an utterance is not a linear process that could be compensated by time scaling but since acceleration of speech rate is obtained by shortening the vowels leaving consonant durations almost unmodified, the comparison requires a non-linear time warping. Dynamic programming is used to find the optimal alignment of test and reference feature vectors and to define a measure of similarity between these two utterances[7]. Comparing all templates with the current test utterance provides the recognized word.

A slight modification of the alignment algorithm allows the recognition of a sentence of concatenated words without inserted pauses: indeed, allowing discontinuities of the optimal path between the end of a word reference and the beginning of a word reference, the best path segments the sentence in words and identifies the uttered words [11].

The major drawbacks of the template approach are:

- The only way to improve the robustness versus intra- and inter-speaker variability is to increase the number of templates by enrolling several references per word and per speaker. The number of models dramatically grows with the size of the lexicon.
- Only templates for words or syllables are possible: indeed phoneme templates are available only at the price of expert work for speech phonetic segmentation. In case of large vocabulary, recognition must definitely rely on phoneme or phoneme-like subunits for speech recognition: f.i. the phoneme inventory of French is less than 40 phonemes from which any word can be represented from its phonetic transcription.

4. HMM-based recognition

A stochastic model of words or even of phonemes is thus proposed [7,12]. A sequence of states q_j connected with transition probabilities (and with self loops) is used to represent a subunit. Each state of this automaton q_i generates the parametric likelihood $p(\vec{x}|q_i)$ that a vector $\vec{x}$ can be associated with it. Parameters of the probability density can be mean vectors, covariance matrices in case of Gaussian distributions and also weights in case of Gaussian mixtures. Vector quantized distributions are also considered where the parameters are the probability of clusters on each state. This automaton is a hidden Markov model (CDHMM for continuous densities and DHMM in case of vector quantization). So, a given utterance is observed as concatenated sequences of vectors

$$X_1^{n_1}, X_{(n_1+1)}^{n_2}, \ldots, X_{(n_L+1)}^{n_{(L+1)}=N}$$

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

generated by successive states of the automaton (respectively $X_{(n_{j-1}+1)}^{n_j}$ on state q_j with $X_p^r \stackrel{\text{def}}{=} \{\vec{x}_p, \ldots, \vec{x}_r\}$). The product of all probabilities and likelihoods met along a path in the automaton is the path probability. The sum over all possible paths is the likelihood that the utterance matches with the current model (in the Baum-Welch training procedure). Frequent use is also made of the likelihood corresponding to the most probable path in the automaton (in Viterbi training and in most of recognizers).

The main characteristic of HMM is that its parameters are trained on examples taken from large databases. Training must be based on the maximization of the a posteriori probability $P(W|X)$ that a given training example X matches with its corresponding model W. This probability is difficult to estimate and using Bayes rule, the criterion turns into the maximization of the likelihood of an example given its model $P(X|W,\lambda)$:

$$\lambda^{opt} = argmax_\lambda P(W|X,\lambda) = argmax_\lambda (P(X|W,\lambda)P(W))$$

where λ denotes the set of parameters of W. Under questionable but usually accepted hypotheses (vector sequence is i.i.d.), likelihood $P(X|W)$ factorizes into a path probability in the model. Probability $P(W)$ is not related to the waveform but depends only on the language model of the application.

Models of sentences are built by concatenating word models according to the word transcription. Viterbi training runs along the following steps:

- Using initial guesses for the parameters of all word models, the best alignment (maximizing $P(X|W)$ corresponding to the best path in the model) between the models and the feature vector stream is searched using dynamic programming.
- This path segments the sentence into states. The partition is used to reestimate the parameters of the models.
- Iteration of this process is convergent and yields trained models.

Baum Welch algorithm [7] also reestimates the parameters iteratively but takes all possible paths in the model into account.

Training provides not only word units but can also provide phoneme models since it is performed on models embedded in sentences whatever the subunits are. Training requires labeled databases (in words or phonemes) but segmentation is obtained from the training algorithm as a by-product.

Observing that the HMM models with Gaussian probability distributions on states are not adequate models for speech signals, distributions have been extended to mixtures of Gaussian distributions. The number of parameters of the models is then growing dramatically.

Another source of increased number of parameters is the use of triphones. Indeed, coarticulation is a major drawback in the use of phonemes i.e. a phoneme is different according to its context: adjacent phonemes severely modify the pronunciation. As a consequence, triphones are used that model

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

phonemes in context. Their training requires very large databases which are usually unbalanced: some triphones occur almost never though they still exist in the lexicon inventory.

Another problem is the lack of discrimination between models. Training increases the probability that an utterance matches with its model (MAP) but ideally this utterance should differ as much as possible from all the other models: this requirement is not taken into account in the MAP training criterion. Different techniques have been proposed to modify the models in order to increase discrimination like Maximum Discriminant Information (MDI)[13], Maximum Mutual Information (MMI) [14] or Corrective Training [15].

The conclusions are:

- Recognition is not a simple classification problem: the decision is made after integration of information over a sequence of feature vectors. For that reason, classification based on statistics is embedded in an automaton (HMM).
- High variability and lack of understanding of the speech production process rule out AI techniques despite their potential flexibility for integrating high level informations like syntax and semantics.
- Training criteria deserve a deeper study to improve discrimination.

Neural networks were considered as promising tools because they are trained classifiers with discriminant properties. Moreover, their outputs will receive a statistical interpretation [16,21].

5. Use of neural networks

5.1. History and tools

The works of J.J. Hopfield [17] and G.Hinton et al.[18] drew the attention of speech scientists on the connectionist machines in the early eighties. In 1986, Prager et al.[1] suggested the use of Boltzmann machines for speech recognition. The idea was interesting but the Boltzmann machine was inadequate due to its high neeed of computation time also in the recognition stage (simulated annealing was required to reach a stable state!). Multilayer perceptrons were also suggested at the IEEE First Annual International Conference on Neural Networks (1987) at San Diego by Bourlard and Wellekens for ASR [2] and by Watrous and Shastri for feature analysis[4]. Also in 1987, Waibel et al. published a report describing the TDNN and its use for isolated phoneme recognition [19]. Later, several tentatives to use radial basis functions (RBF) were also published. Their role was equivalent to what MLP play in hybrid networks. They constitute an alternative to mixtures of Gaussian distributions used in HMM.

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

5.2. MLP

The individual classification of feature vectors into phonetic classes is only loosely related to speech recognition. However, work on this task puts into the light the fundamental role of context and the stochastic interpretation of the output values of a MLP. Indeed, a first set of experiments demonstrated that a MLP with a single feature vector in its input field and with as many outputs as possible phonemic classes generates in its output field the probabilities that the input vector belongs to each phonemic class $p(q_i|\vec{x})$, if it has been trained for desired output values 1 or 0 according to current phoneme. It is even easy to check that the outputs sum up to 1 as it can be expected from this interpretation [16,21]. The discrimination between classes is also enhanced if the input field is enlarged with right and left contexts of the current feature vector. Once this property has been recognized, tentatives were made to use these local a posteriori probabilities rather than the state density functions for continuous speech recognition as explained in the next section.

Phoneme classification has also been achieved by Waibel [19] using a TDNN (this is an MLP architecture with memories in each layer connected to the next one). There is some effect of integration due to the in-layer memories but continuous speech recognition is not possible without an HMM. Another attempt is due to Robinson and Fallside [20] where use is made of a recurrent MLP where internal states are reinjected with a delay in the extended input field. This recurrence is an alternative way to inject context dependency in the input signal. An interesting approach for discriminant classification is proposed by Juang and Katagiri[22]. They define a new misclassification measure based on classification functions. Following the neural network approach, this measure is squashed with a sigmoid and they minimize the risk of misclassification described in terms of squashed classification functions. This leads to an MLP architecture trained with the minimum classification risk criterion.

Vector quantization is also currently used for classification and its training is unsupervised. Other unsupervised selforganizing mappings have been used for preclassification into phonetic classes [25].

Learning vector quantization (LVQ) is a classifier trained in a supervised way and increasing discrimination by competitive learning. It has also been used for classification [24] in conjunction with HMM to form an hybrid network.

6. Hybrid networks

An excellent tutorial on hybrid HMM/ANN is by Morgan and Bourlard [23].

Discovering that MLP outputs can be considered as estimates of local a posteriori probabilities lead researchers to use these probabilities instead of the densities associated with the states but this approach had no robust theoretical justification and led of course to disappointing results. Use of Bayes rule

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

transforms this local a posteriori probability into emission likelihood $p(\vec{x}|q_i)$:

$$p(\vec{x}|q_i) = \frac{p(q_i|\vec{x})p(\vec{x})}{p(q_i)}.$$

So, dividing the output of a MLP by the a priori probability $p(q_i)$ gives an estimate of $p(\vec{x}|q_i)$ within a factor $p(\vec{x})$ irrelevant for path building. This estimate of the emission likelihood is a result of discriminant training and if used in conjunction with a HMM, it gives excellent scores even on large databases (TIMIT, RM, WSJ,...). In the preliminary tests, models for phonemes were single states and the database had to be segmented and priors were estimated from the database by simple counting but as soon as MLP training was embedded in the Viterbi training, classical 3-state phoneme models were used and only labeling in phonemes of the data base was required [27].

An interesting approach is based on the consideration that the recognition criterion is the a posteriori probability which is systematically circumvented in most of recognition algorithms. A new idea was to formulate the problem in terms of local conditional transition probabilities: this leads to a direct use of the a posteriori probability hence the name REMAP. Training follows a modified Baum Welch algorithm and a very interesting result is that updating iteratively the desired outputs of the MLP increases the probability and guarantees convergence [28].

A completely different way to use MLP in hybrid networks is based on the observation that taking contextual aspects of a feature vector into account is equivalent to use its prediction error as a local distance [29]. E.Levin suggested use non-linear pridiction error where the MLP is used to predict the current vector from several previous ones [30]. An approach based on similar ideas was presented by Tebelskis and Waibel [31]. A weakness of non-linear predictors is the lack of discrimination since the MLP is no longer used as a discriminant classifier.

7. Conclusions

Connectionist approach gave ASR a new blood in the eighties. It opened new roads and elicit new investigations even in the traditional domain of HMM for ASR. Rediscussion of the criteria, sophistication of the probability densities to better fit with the underlaying reality as the exceptional mapping properties of neural networks do, contextual processing of the information have to be credited to this new point of view on the problems. Neural networks have also been used in many other applications in speech processing than ASR.

Dedicated architectures and chips for training neural networks have been created and it is not overclaiming to say that speech processing has supported the general research in connectionism. It still will do in the future.

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

References

[1] R.W. Prager, T.D.Harrison, F.Fallside, "Boltzmann machines for speech recognition," *Computer, Speech and Language*, 1, pp 3-27, Academic Press,1986

[2] H. Bourlard, C.J. Wellekens, "Multilayer Perceptrons and Automatic Speech Recognition," *IEEE First Int. Conf. on Neural Networks*, San Diego, Calif. IV-407-IV-416, June 21-24,1987

[3] H.Bourlard, C.J.Wellekens, "Speech pattern discriminations and multilayer perceptrons," *Computer, Speech and Language*, Dec 1987

[4] R.L.Watrous,L. Shastri, "Learning phonetic features using connectionist networks," *IEEE First Int. Conf. on Neural Networks*, San Diego, Calif. IV-381-388, June 21-24,1987

[5] H.Bourlard, N.Morgan,*Connectionist Speech Recognition- A Hybrid Approach*, Kluwer Academic Press, 1994.

[6] M.Hochberg, S.Renals, A.Robinson, "ABBOT: the CUED Hybrid Connectionist-HMM Large Vocabulary Recognition System," *Proc. Spoken Language Technology Workshop*, pp. 170-178, Morgan Kaufmann Publishers Inc, Austin, Texas, Jan 1995.

[7] L.R. Rabiner and B.H. Juang, *Fundamentals in Speech recognition*, Prentice Hall, 1993.

[8] J.R.Deller, J.G.Proakis, J.H.L.Hansen, *Discrete Time Processing of Speech Signals*, MacMillan Publishing Company,1993

[9] J.Cl. Junqua and J-P Haton, *Robustness in Automatic Speech Recognition, Fundamentals and Applications*, Kluwer Academic Publishers, 1996

[10] S.Furui, "Speaker independent isolated word recognizer using dynamic features of speech spectrum," *IEEE Trans. ASSP*,,vol.34, nr 1, pp.52-59, 1986.

[11] H. Bourlard, Y.Kamp, H.Ney, C.J.Wellekens, "Speaker-Dependent Connected Speech recognition via Dynamic Programming and Statistical Methods," in *Speech and Speaker Recognition*, Ed. M.R. Schroeder, Karger,1985.

[12] F.Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. of the IEEE*, vol.64, nr 4,pp. 532-555, 1976.

[13] Y.Ephraim, A.Dembo, L.R. Rabiner, "A minimum discrimination information approach for hidden Markov models," *IEEE Transactions on Information Theory,*vol. 35, pp.1001-1013, September 1989

[14] L.R.Bahl, P.F.Brown, P.V.DeSouza et al, "Maximum Mutual Information estimation of hidden Markov model parameters for speech recognition," *Proc. ICASSP-86*, vol.1 pp.49-52, Tokyo

[15] L.R.Bahl, P.F.Brown, P.V.DeSouza et al, "A new algorithm for the estimation of hidden Markov model parameters,"*Proc.ICASSP-88* vol.1, pp. 493-496, New York.

[16] H. Bourlard and C.J.Wellekens, "Links between Markov Models and multilayer perceptrons," *IEEE Trans. on PAMI,*vol 12,pp.1167-1178, 1990.

[17] J.J.Hopfield,"Neural networks and physical systems with emergent computational properties," *Proc. of Nat. Academy of Science, USA*, 81,pp.3088-3093, 1982

[18] G.E.Hinton, T.Sejnowski and D.H. Ackley, "Boltzmann machines: constraint satisfaction networks that learn," *Techn.Report. CMU-CS-84-119, Carnegie Mellon University*,1984.

[19] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, K.Lang, Phoneme recognition Using Time Delay Neural Networks. *Technical report, ATR Interpreting telephony Research Laboratory*, Kyoto,1987.

[20] A.Robinson and F.Fallside, "A Recurrent Error Propagation Network Speech Recognizer System," *Computer, Speech and Language*, vol.5, nr 3, July 1991.

[21] M.D. Richard and R.Lippman, "Neural network classifiers estimate Bayesian a posteriori probabilities," *Neural Computation*, nr 3, pp.461-483, 1991.

[22] B.H. Juang and S.Katagiri, "Discriminative Learning for Minimum Error Classification," *IEEE Trans. Signal Processing*, vol.40, nr 12, pp.3043-3054, Dec 1992

[23] N.Morgan and H.Bourlard, "Continuous Speech recognition, An introduction to the hybrid HMM/connectionist approach," *IEEE Signal Processing Magazine*, vol. 12, nr 3, pp.25-42, May 1995.

[24] T.Kohonen, "An introduction to neural computing", *Neural Networks*,vol 1, pp3-16, 1988.

[25] T.Kohonen, "The neural phonetic typewriter," *IEEE Computer*, 11-2, 1988.

[26] T.Robinson, L.Almeida, J.M.Boite, H.Bourlard, F.Fallside, M.Hochberg, D.Kershaw, P.Kohn, Y.Konig, N.Morgan, J.P. Neto, S.Renals, M.Saerens, C.Wooters, "A neural network based, speaker independent, large vocabulary, continuous speech recognition system: The WERNICKE Project," *Proc. EUROSPEECH 93*, Berlin, pp.1941-1944.

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 227-236

[27] H.Bourlard and N.Morgan," A continuous speech recognition system embedding MLP into HMM," in *Advances in Neural Information Processing Systems 2*, (D.S.Touretzky, Ed.), pp.413-416, Morgan Kaufmann, San Mateo CA.,1990.

[28] H.Bourlard, Y.Konig, N.Morgan, "REMAP: recursive estimation and maximization of a posteriori probabilities: Application to transition based connectionist speech recognition," *ICSI Technical Report TR-94-064*, 1994

[29] C.J. Wellekens, "Explicit Time Correlation in Hidden Markov Models for Speech Recognition," *Proc. IEEE Conference on Acoustics, Speech & Signal Processing ICASSP-87*, vol. 1, pp. 384-386, Dallas, April 1987.

[30] E.Levin, " Speech recognition using hidden control neural network architecture," *Proc. ICASSP-90*, pp. 433-436, Albuquerque (NM), 1990

[31] J.Tebelskis, A.Waibel, "Large vocabulary recognition using linked predictive neural networks," *Proc. ICASSP-90*, pp.437-440, Albuquerque (NM), 1990