training, however, is not a common approach for building a statistical discrimination function. A method of including ambiguous samples for network training is currently under investigation [15].

In conclusion, a procedure was developed for making voiced, unvoiced, and silence classifications of speech using an MFN. The network V/U/S classifier is expected to provide a useful tool for speech analysis and may also have applications in speech-data mixed communication systems.

References

[1] B. Atal and S. Hanuer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Amer.*, vol. 50, pp. 637–655, Aug. 1971.

[2] B. Atal and L. Rabiner, "A pattern recognition approach to Voiced-Unvoiced-Silence classification with applications to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201–212, June 1976.

[3] L. Siegel, "A procedure for using pattern classification techniques to obtain a Voiced/Unvoiced classifier," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 83–88, Feb. 1979.

[4] R. Duda and P. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[5] L. Siegel and A. Bessey, "Voiced/Unvoiced/Mixed excitation classification of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 451–460, June 1982.

[6] L. Rabiner and M. Sambur, "Application of an LPC distance measure to the Voiced-Unvoiced-Silence detection problem," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 338–343, Aug. 1977.

[7] D. Rumelhart, G. Hinton, and R. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructures of Cognition* D. Rumelhart and J. McClelland, Eds., vol. 1,\quad Cambridge, MA: MIT Press, 1986, pp. 318–362.

[8] G. Fant, "The source filter concept in voice production," *QPSR—Speech Transmission Laboratory*, vol. 1, pp. 21–37, 1981.

[9] R. Lippman, "An introduction to computing with neural nets," *IEEE ASSP Mag.*, vol. 1, pp. 4–22, 1987.

[10] D. Ruck, S. Rogers, M. Kabrisky, M. Oxley, and B. Suter, "The multilayer perceptron as an approximation to a Bayes optimal discriminant function," *IEEE. Trans. Neural Networks*, vol. pp. 296–268, Dec. 1990.

[11] A. Gray and J. Markel, "Distance measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 381–391, Oct. 1976.

[12] Chang and Fallside, "An adaptive training algorithm for bp networks," *Computer Speech and Language*, pp. 205–218, 1987.

[13] L. Niles, H. Silverman, G. Tajchman, and M. Bush, "How limited training data can allow a neural network to outperform an optimal statistical classifier," in *Proc. ICASSP89*, vol. 1, pp. 17–20, 1989.

[14] L. Niles, H. Silverman, G. Tajchman, and M. Bush, "The effects of training set size on relative performance of neural network and other pattern classifiers," *Tech. Rep. LEMS-51*, Brown University, Providence, RI, 1989.

[15] B. Hunt, Y. Qi, and D. Dekruger, "Fuzzy classification using set membership functions in the back propagation algorithm," *Heuristics, J. Knowledge Eng.*, vo. 5, no. 2, pp. 62–74, 1992.

On the Locality of the Forward–Backward Algorithm

Bernard Merialdo

***Abstract*—In this paper, we present a theorem which shows that the local maximum found by the Forward–Backward algorithm in the case of discrete hidden Markov models is really "local." By this we mean that this local maximum is restricted to lie in the same connected component of the set $\{x : P(x) \geq P(x_0)\}$ as the initial point x_0 (where $P(x)$ is the polynomial being maximized). This theoretical result suggests that, in practice, the choice of the initial point is important for the quality of the maximum obtained by the algorithm.**

I. Introduction

Hidden Markov models are increasingly being used in various domains and, in particular, in speech recognition [1], [7]–[9]. Their popularity comes from the existence of an efficient training procedure, which, given an observed output string, allows the values of their parameters (transition and emission probabilities) to be estimated. This procedure is known as the *Baum–Welch algorithm* or the *Forward–Backward algorithm.* It is an iterative algorithm which starts from an initial point (a set of parameter values) and builds a sequence of reestimates which improve the likelihood of the training data. This sequence converges to a local maximum of the likelihood function.

A detailed presentation of the theory and practice of hidden Markov models can be found in [11]. Nadas [10] discusses the use of the Baum–Welch algorithm and makes some remarks on the choice of the initial point.

II. The Baum–Welch Algorithm

In the discrete case (i.e., when the output symbols belong to a finite alphabet), the convergence of this algorithm comes from the following theorem:

Theorem A [3], [4]: Let $p\ (X) = p\ (\{X_{ij}\})$ be a polynomial with positive coefficients, homogeneous of degree d in its variables X_{ij}.

Let $x = \{x_{ij}\}$ be any point of the domain:

$$D : x_{ij} \geq 0,\ \sum_{j=1}^{q_i} xij = 1, \qquad j = 1, \cdots, q_i$$

such that,

$$\sum_{j=1}^{q_i} x_{ij} \frac{\partial P}{\partial X_{ij}}(x) \neq 0, \qquad \text{for all } i.$$

Let $y = T_p(x)$ denote the point defined by

$$y_{ij} = \left(x_{ij} \frac{\partial P}{\partial X_{ij}}(x)\right) \Big/ \left(\sum_{j=1}^{q_i} x_{ij} \frac{\partial P}{\partial X_{ij}}(x)\right).$$

Then,

$$P(T_p(x)) > P(x) \ \text{ unless } T_P(x) = x.$$

From Theorem A we can see that when we choose an initial point x_0 and build the sequence of iterates:

$$x_{i+1} = T_P(x_i)$$

Manuscript received June 6, 1991; revised July 6, 1992. The associate editor coordinating the review of this paper and approving it for publication is Dr. Brian A. Hanson.

The author is with IBM France Scientific Center, 75001 Paris, France.

IEEE Log Number 9206396.

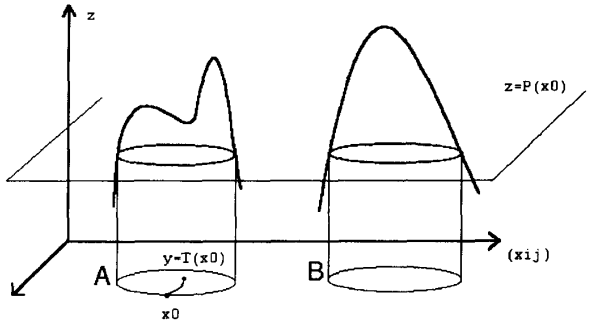

Fig. 1. Connexity and iterations of the BW algorithm.

we always improve the value of $P(x)$. Other simple topological arguments ensure that the sequence of iterates will converge to a limit point, which will be a local maximum of $P(x)$.

III. Locality of the Maximum

As it stands, the demonstration of Theorem A only guarantees that the algorithm will converge to a local maximum. Hopefully, and in the absence of other evidence, this local maximum might be good enough, and may even produce the global maximum. In Fig. 1, for example, if the starting point is x_0, the algorithm would have to "jump" from region A to region B to get the higher maximum.

However, it has often been found experimentally that the choice of the initial point has some influence on the quality of the maximum. For example, to train an acoustic model for a new speaker it is often better to take the model of another speaker as initial point rather than to start from uniform statistics.

We are going to show that there are theoretical limitations for the Forward–Backward algorithm and that the choice of initial point severely restricts the set of local maxima that can be reached. Our result is based on the following theorem.

Theorem B: The point $T_P(x_0)$ lies in the same connected component of the set $\{x : P(x) \geq P(x_0)\}$ as the initial point x_0.

Proof: The proof is directly inspired by some arguments introduced in [6]. Consider the polynomial:

$$R(X) = \sum_{i=1}^{P} \left(\sum_{j=1}^{q_i} X_{ij} \right)^d.$$

This polynomial is homogeneous of degree d with positive coefficients and is constant over domain D. Now we introduce the polynomial:

$$Q(X) = P(X) + \lambda R(X)$$

where λ is a positive constant. We can apply Theorem B to the polynomials $P(X)$ and $Q(X)$. Let

$$y = T_P(x_0) \text{ and } z = T_Q(x_0).$$

Since the derivative of $Q(X)$ is

$$\frac{\partial Q}{\partial X_{ij}}(X) = \frac{\partial P}{\partial X_{ij}}(X) + \lambda d \left(\sum_{k=1}^{q_i} X_{ik} \right)^{d-1}$$

we have

$$z_{ij} = \frac{x_{0ij}\left(\frac{\partial P}{\partial X_{ij}}(x_0) + \lambda d\right)}{\sum_{j=1}^{q_i} x_{0ij}\left(\frac{\partial P}{\partial X_{ij}}(x_0) + \lambda d\right)} = \frac{x_{0ij}\frac{\partial P}{\partial X_{ij}}(x_0) + \lambda d x_{ij}}{\left(\sum_{j=1}^{q_i} x_{ij}\frac{\partial P}{\partial X_{ij}}(x_0)\right) + \lambda d}.$$

Let us introduce the notation:

$$C_{i,P} = \sum_{j=1}^{q_i} x_{0ij} \frac{\partial P}{\partial X_{ij}}(x_0).$$

(Note that $C_{i,P}$ does not depend on λ and is positive.) Now we have

$$z_{ij} = \frac{C_{i,P}}{C_{i,P} + \lambda d} y_{ij} + \frac{\lambda d}{C_{i,P} + \lambda d} x_{0ij}.$$

Therefore, we have

$$Q(z) \geq Q(x_0)$$

and since $R(z) = R(x_0)$, this implies

$$P(z) \geq P(x_0)$$

. Let us note z_λ to recall the dependency of z on λ. We have just proved that, for all λ:

$$P(z_\lambda) \geq P(x_0).$$

When λ varies from infinity down to 0, z_λ follows a curve from x_0 to y. Along this curve we always have $P(z_\lambda) \geq P(x_0)$. Therefore, y is in the same connected component of the set $\{x : P(x) \geq P(x_0)\}$ as the initial point x_0.

We have shown that the first iterate lies in this connected component. By transitivity, all subsequent iterates will also be in the same component, as will be the local maximum which is their limit. Taking the example of Fig. 1, Theorem B indicates that it is in fact impossible to "jump" to region B, and that the algorithm can only find one of the two local maxima of region A.

One should argue that, by taking a bad initial point with a low value for $P(x_0)$, the connected component could be sufficiently vast to include the best maximum with high probability. Note however, that Theorem B can be applied to any step of the iteration so that as soon as an iterate point starts climbing a hill, the following iterates have to climb the same hill and cannot jump to another one. Therefore, if the iteration process reaches a point on the hill, it cannot reach a better maximum than the local maximum corresponding to this hill.

IV. Conclusion

We have presented a theorem which shows that in the case of discrete hidden Markov models, the Baum–Welch (or Forward–Backward) algorithm can only find a local maximum that is closely related to the position of the initial point. In particular, once it starts going up a given hill, it can never go down through a valley to reach another hill.

This gives some theoretical support for the use of a good starting point obtained by some preliminary estimation, rather than by random or uniform statistics.

Acknowledgment

As pointed judiciously by a reviewer, Theorem B was already known in the 1960's, though using a different proof. For example, a somewhat stronger statement was proved by J. D. Ferguson in his book in *Variable Duration Models for Speech* [5].

REFERENCES

[1] A. Averbuch *et al.*, "Experiments with Tangora 20,000 word speech recognizer," in *ICASSP*, Dallas, TX, pp. 701–704, 1987.

[2] L. R. Bahl, F. Jelinek, and R. L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. PAMI-5, Mar. 1983.

[3] L. E. Baum, "An inequality and associated maximization technique in statistical estomation for probabilistic function of Markov processes," *Inequality*, vol. III, pp. 1–8, 1972.

[4] L. E. Baum and J. A. Eagon, "An inequality with application to statistical estimation for probabilistic function of Markov processes and to a model for ecology," *Bull. AMS*, vol. 73, pp. 360–363, 1967.

[5] J. D. Ferguson, *Hidden Markov Models for Speech Recognition*. Princeton, NJ: IDA/CRD, 1980.

[6] P. S. Gopalakrishnan, D. Kanevsky, A. Nadas, and D. Nahamoo, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Trans. on Inform. Theory*, vol. 37, pp. 107–117, Jan. 1991.

[7] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, Apr. 1976.

[8] K. F. Lee, H. W. Hon, and R. Reddy, "An overview of the SPHINX speech recognition system," *IEEE Trans. Acoust., Speech., Signal Processing*, vol. 38, pp. 35–45, Jan. 1990.

[9] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application od the theory of probabilistic functions of a Markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, pp. 1035–1074, 1983.

[10] A. Nadas, "Hidden Markov chains, the forward–backward algorithm and initial statics," *IEEE Trans. Acoust., Speech., Signal Processing*, vol. ASSP-31, pp. 504–506, 1983.

[11] A. B. Poritz, "Hidden Markov models: A guided tour," in *ICASSP*, New York, pp. 7–13, Apr. 1988.

Convergence of Acoustic Echo Cancellers for Hands-Free Telephones Operating Under Feedback Conditions

H. Schütze

***Abstract*—Acoustic echo cancellers for conference circuits with hands-free telephones are capable of identifying a time-invariant room without problems, if no disturbances appear and no feedback exists over a second room. In practice, these conditions are rarely fulfilled. In the following paper, the conditions are given, under which acoustic echo cancellers within a closed-loop conference circuit attain the same convergence behavior as in the open-loop case without additional difficulties.**

I. INTRODUCTION

An acoustic echo canceller (AEC) has to identify the loudspeaker-room-microphone (LRM) system at its subscriber end to be able to permanently suppress its echo signal (Fig. 1). For the identification algorithms whose convergence characteristics are well known (e.g., [1], [2], the convergence statements are usually applicable for the case without feedback. However, conference circuits always represent feedback systems which, additionally, are time-variant and whose identification process is disturbed by signals of their own subscriber station. It is the objective of this paper to show under which conditions the known convergence statements for AEC's in open loops also are valid in closed ones.

Manuscript received June 14, 1991; revised February 28, 1992. The associate editor coordinating the review of this paper and approving it for publication was Dr. Sharad Singhal.

The author is with Deutsche Bundepost TELEKOM Research Institute, 1000 Berlin 42, Germany.

IEEE Log Number 9206398.

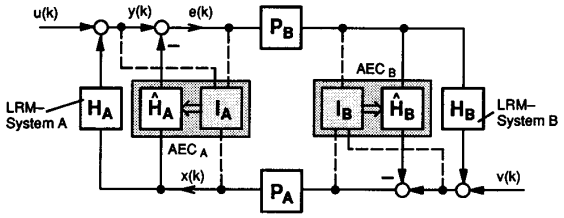

Fig. 1. Conference circuit comprising two loudspeaker–room–microphone (LRM) systems with acoustic echo cancellers (AEC).

II. SYSTEM DESCRIPTION

A hands-free conference circuit with an AEC on either side has the basic structure shown in Fig. 1. The two participating stations are named A and B. All following considerations refer always to the identification of LRM system A. For LRM system B our statements apply analogously.

A. LRM System

The LRM system A to be identified is assumed to be linear, time-invariant, and time-discrete:

$$\boxed{y(k) = H_A(q^{-1})x(k) + u(k).} \tag{1}$$

Here $x(k)$ is an exciting input signal, $u(k)$ is a disturbing voice signal or background noise from room A. The LRM system A is described by an FIR filter,

$$H_A(q^{-1}) = \sum_{i=0}^{n_A} h_{A,i}q^{-i};\quad H_A(0) = h_{A,0} = 0. \tag{2}$$

where q^{-i} is a shift operator with the effect $q^{-i}x(k) = x(k-i)$.

B. Feedback System

The feedback comprises all remaining transmission systems in Fig. 1 except $H_A(q^{-1})$.

- $\hat{H}_A(k, q^{-1})$ is a model of $H_A(q^{-1})$ which is set by the identification algorithm I_A:

$$\hat{H}_A(k, q^{-1}) = \sum_{i=0}^{n_A} \hat{h}_{A,i}(k)q^{-i};\quad \hat{H}_A(0) = \hat{h}_{A,0}(k) = 0 \tag{3}$$

- $H_B(k, q^{-1})$describes LRM system B:

$$H_B(k, q^{-1}) = \sum_{i=0}^{n_B} h_{B,i}(k)q^{-i};\quad H_B(0) = h_{B,0}(k) = 0 \tag{4}$$

- $\hat{H}_B(k, q^{-1})$ is a model of $H_B(k, q^{-1})$ which is set by identification algorithm I_B:

$$\hat{H}_B(k, q^{-1}) = \sum_{i=0}^{n_B} \hat{h}_{B,i}(k)q^{-i};\quad \hat{H}_B(0) = \hat{h}_{B,0}(k) = 0 \tag{5}$$