Automatic Speech Segmentation using Neural Network and Phonetic Transcription

Harald Finster

Institute for Communication Systems and Data Processing

Aachen University of Technology

Templergraben 55

5100 Aachen, Germany

Introduction:

Speech segmentation is a basic task to train continuous speech recognizers or speech synthesizers. As manual segmentation is extremely time consuming and depending on subjective criteria of phoneticians there is a great interest in the development of automatic speech segmentation algorithms [1].

Known segmentation algorithms are based on the stationarity of phoneme segments [1], make pre-assumptions about phoneme-durations [5] or make use of sophisticated phonetical rules [3].

In this contribution a new algorithm for automatic segmentation of speech based on its phonetic transcription is proposed.

The specific features are:

- new iterative self-learning procedure to find the temporal alignment between feature vectors and phonetic transcription
- no preassumptions about statistical speech properties or phonetical rules
- no pretraining required !

System-description :

The general structure of the segmentation system is shown in Fig. 1.

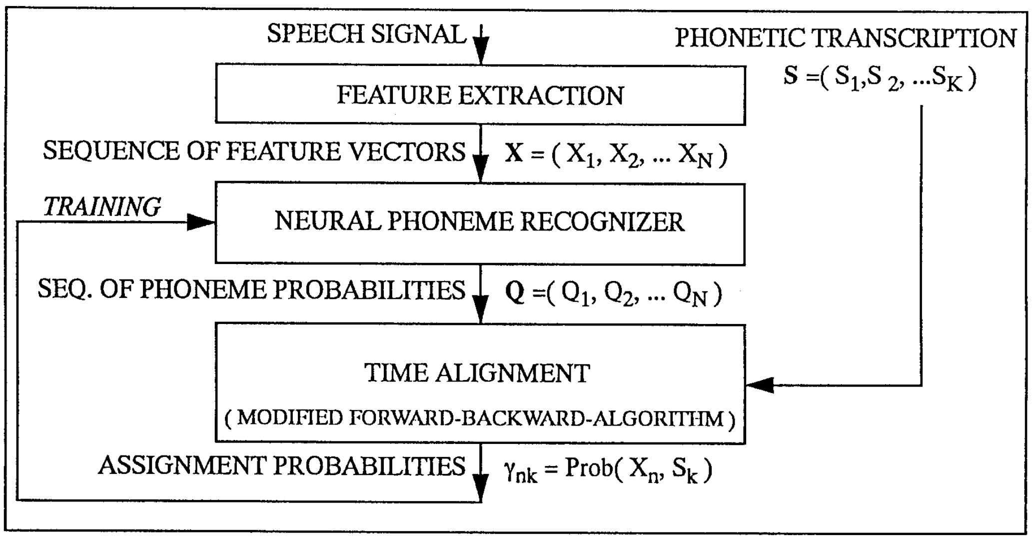

Fig. 1. General structure of the segmentation system

In the first step a sequence of feature vectors **X** is generated e.g. by a 19 channel mel scale filter bank [7] with one vector X_n each 50 ms

$$\mathbf{X} = (X_0, X_1, \ldots X_n, \ldots X_N) \qquad (n = \text{time index})$$

0-7803-0559-0 /92 $3.00

representing the speech signal.

The core of the segmentation procedure is an iterative loop consisting of a neural phoneme classifier, a time-alignment algorithm and the retraining of the neural classifier.

The neural net, a three-layer perceptron [4], calculates the probabilities q_{ni} of the different phonemes S_i for each incoming feature vector X_n. The probability, that the feature vector X_n represents the phoneme S_i is denoted by

$$q_{ni} = \mathrm{Prob}(S_i \mid X_n) \qquad \text{with} \quad S_i \in \{ \text{'\_'}, \text{'a'}, \text{'e'}, \dots \text{'z'} \}$$

If L is the number of different phonemes the phoneme probability vector is given by

$$Q_n = (q_{n1}, q_{n2}, \dots q_{ni}, \dots q_{nL}) \qquad (\, n = \text{time index},\ i = \text{symbol index}\,)$$

The analysis of the entire speech sequence results in the sequence of phoneme probability vectors:

$$Q = (\, Q_1, Q_2, \dots Q_n, \dots Q_N\,)$$

In the second step of the iteration, assignment probabilities between the sequence of phoneme-probabilities **Q** and the known phonetical transcription

$$S = (S_1, S_2, \dots S_k, \dots S_K) \qquad \text{with} \quad S_k \in \{ \text{'\_'}, \text{'a'}, \text{'e'}, \dots \text{'z'} \}$$

are calculated by a modified forward-backward algorithm [6].

$$\gamma_{nk} = \mathrm{Prob}(X_n, S_k \mid X, S)$$

γ_{nk} is the probability, that the feature vector X_n represents the k-th symbol S_k of the phonetic transcription S in time n, given the sequence of feature vectors **X** and the phonetic transcription **S**.

Finally the network is re-trained by error back-propagation [4] according to the assignment probabilities γ_{nk}. For each time n=1 ... N a feature vector X_n is fed into the net as input. The network is trained with each of the phonemes S_k (k = 1 ... K) as the desired output with learning rate γ_{nk}.

In this way assignments between features and phonemes, which are securely right, will be learned relatively faster than those, which are assigned with lower probability.

After a sufficient number of iterations the matrix of the assignment probabilities γ_{nk} gives the optimal assignment path between the feature vectors and the phonetical transcription.

Results:

Fig. 2 shows the the segmentation of the sentence 'nine two seven eight nine ten'.

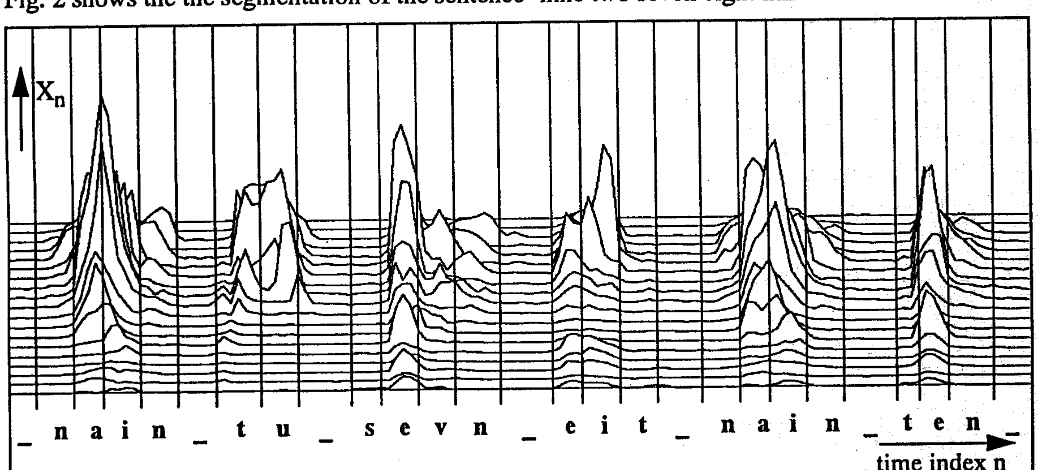

Fig. 2. Segmentation of the sentence 'nine two seven eight nine ten'

It can be seen, that all boundaries were found correctly. The algorithm produces neither deletions nor insertions of phonemes.

Fig.3 shows the initial (3a) and the final (3b) assignment probabilities found by the alignment procedure.

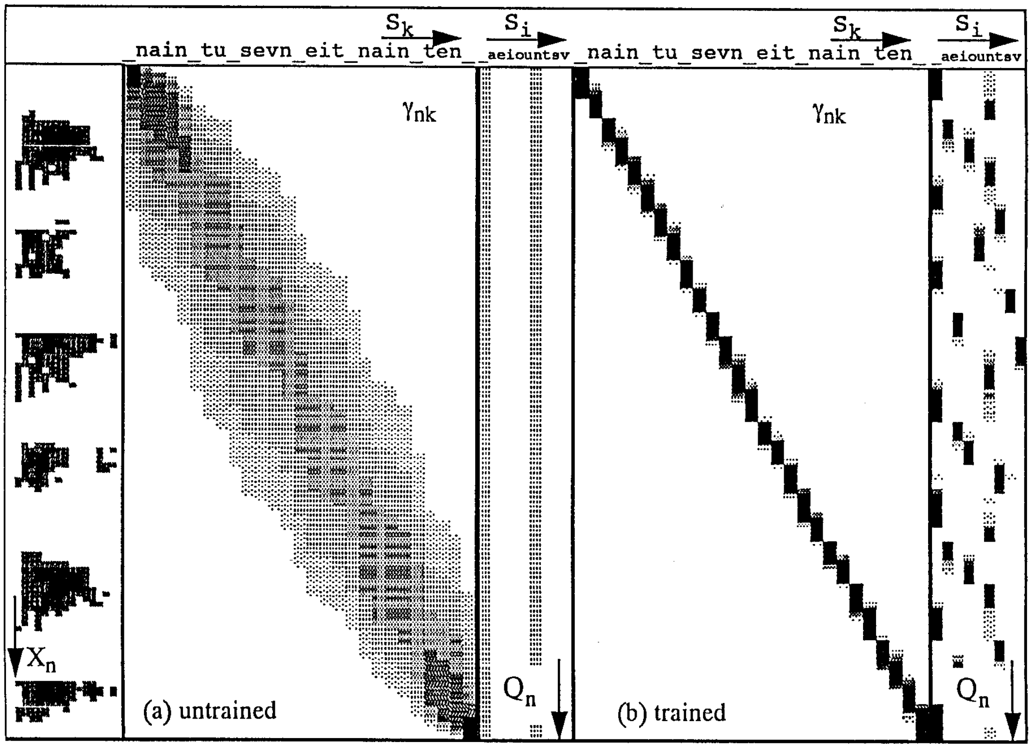

Fig. 3. Example: assignment probabilities between feature vectors and phonemes ($S_i \in \{$ '_', 'a', 'e', 'i', 'o', 'u', 'n', 't', 's', 'v' $\}$)

References :

[1] J. P. van Hemert, "Automatic Segmentation of Speech", IEEE TRANSACTIONS ON SIGNAL PROCESSING, Vol. 39, NO. 4, April 1991.

[2] S. Imai, C. Furuichi, "Automatic labeling of continuous japanese speech into phonetic units", Signal Processing V: Theories and Applications, L. Torres, E. Masgrau, M. A. Lagunas (eds.) Elsevier Science Publishers B. V., 1990.

[3] H. C. Leung, V. W. Zue, "A procedure for automatic alignment of phonetic transcriptions with continuous speech", in Proc. 1984 IEEE Int. Conf. Acoust. Speech, Signal Processing, 1984 pp. 2.7.1-2.7.4.

[4] R.P. Lippman, "An Introdiuction to Computing with Neural nets", IEEE ASSP Magazine, April 1987

[5] D. Mergel, H. Ney, "Phonetically guided clustering for isolated word recognition", in Proc. 1985 IEEE Int. Conf. Acoust. Speech, Signal Processing, 1985, pp. 854-857.

[6] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, VOL. 77, NO. 2, Feb. 1989

[7] E. Zwicker, Psychoakustik (Hochschultext), Springer Verlag, 1982