S3.8

An Improved Sub-Word Based Speech Recognizer

T. Svendsen, K. K. Paliwal(*), E. Harborg, P. O. Husøy(*)

Div. of Telecommunications and (*)ELAB
The Norwegian Institute of Technology
N-7034 Trondheim
Norway

Abstract: Whole-word based speech recognition has proven successful for the recognition of small and medium-sized vocabularies. For large vocabularies and/or continuous speech, the use of sub-word reference units is a promising and efficient alternative. We describe a system for speaker dependent speech recognition based on acoustic sub-word units. Several strategies for automatic generation of an acoustic lexicon are outlined. Preliminary tests have been performed on small vocabulary. In these tests, the proposed system showed comparable results to whole-word based systems.

1. Introduction

Although the research in the area of automatic speech recognition has been pursued for the last three decades, only whole-word based speech recognition systems have found practical use and have become commercial successes [1-4]. Two important reasons for this success are that the effects of context-dependence and coarticulation within the word are implicitly built into the word models and that there is no necessity of lexical decoding.

In spite of their success, these whole-word based speech recognition systems suffer from two problems:

1) Coarticulation effects across the word boundaries. This problem has been reasonably well solved and connected word recognition systems with good performance have been reported in the literature [1], [4].

2) Amount of training data. It is extremely difficult to obtain good whole word reference models from a limited amount of speech data available for training. This training problem becomes even worse for large vocabulary speech recognition systems.

In order to overcome these problems, a sub-word based approach is a viable alternative to the whole-word based approach. Here, the word models are built from a small inventory of sub-word units.

Traditionally, the definition of sub-word units employed in speech recognition has been defined based upon a linguistic description of the language. Typical examples are phonemes, diphones and demi-syllables. This approach has the advantage that the lexical decoding is an easily accomplished task as word models based on these units already exist. There is, however, a major problem when it comes to correctly detecting and identifying these units. This is due to the mismatch between the acoustically based analysis of the actual speech signal and the linguistically based description of the language. Because of this mismatch, the performance of the speech recognition system based on linguistic sub-word units is inferior to that of whole-word based systems[6].

Recently, acoustically defined sub-words have been used as units in speech recognition systems[7-11]. These units need not have a one-to-one correspondence to any linguistic units. Segmentation of the speech utterance in terms of these units can be done using well defined acoustic criteria. Thus, there is no mismatch problem as encountered with the linguistic sub-word units and hence, performance of the recognition systems based on these units can be expected to be as good as that of the whole-word based speech recognition systems. Since it is not (yet) possible to attach a linguistic interpretation to these units, the lexical decoding is no more a straightforward task. However, it is conceivable to develop automatic procedures for generating word lexica based on acoustically defined sub-word units.

In the present paper, we study the use of acoustic sub-word models for speech recognition. We consider the recognition system of Lee et al.[9] as a base-line system.

In the training phase, our system requires the following operations:

1) Segmentation of speech utterances into acoustic segments.
2) Clustering of the acoustic segments into N clusters where N is the prefixed number of sub-word units used in our system.
3) Labelling of the acoustic segments into sub-word classes and forming sub-word clusters.
4) Generation of HMM model for each sub-word from the acoustic segments in its cluster.
5) Generation of acoustic lexicon.

In the recognition phase, the input utterance is compared with the different word models where each word model is generated as a sequence of the sub-word units using the acoustic lexicon and the recognition is performed using the maximum likelihood decision rule.

2. Maximum likelihood segmentation

The aim for the algorithm is to segment the speech signal in such a way that the frames contained in each segment are acoustically similar. An automatic method for performing this segmentation is maximum likelihood segmentation[13].

We assume that a frame of K speech samples, $\boldsymbol{x}_n=[x_n(1),x_n(2),....,x_n(K)]^T$ at time n can be viewed as a realization of a stationary, Gaussian AR(p) process with an approximate pdf

$$p(\boldsymbol{x}_n \mid \sigma_n,\boldsymbol{a}_n)\sim(2\pi\sigma_n^2)^{-(K/2)}\exp\{-(\boldsymbol{a}_n^T\boldsymbol{R}_n\boldsymbol{a}_n)/2\sigma_n^2\} \quad (1)$$

where $\boldsymbol{a}_n=[1,a_1,....,a_p]^T$ is the LPC inverse filter coefficients, σ_n^2 is the variance of the innovations process and $\boldsymbol{R}_n$ is the pxp autocorrelation matrix of $\boldsymbol{x}_n$. Assuming that T successive frames of speech are generated by the same AR(p) model, their joint pdf is the product of their respective pdf's.

If we, for a given speech utterance $\{\boldsymbol{x}_1,\boldsymbol{x}_2,....,\boldsymbol{x}_T\}$ wish to segment the utterance into m consecutive segments with segment boundaries $\{b_1,b_2,....,b_m\}$ such that the overall likelihood function is maximized, this is equivalent to minimizing the overall likelihood ratio distortion

$$\min_{\{b_1,b_2,...,b_m\}} \sum_{i=1}^{m}\sum_{n=b_i+1}^{b_{i+1}} (\boldsymbol{a}_i^T\boldsymbol{R}_n\boldsymbol{a}_i)/\sigma_i^2 \quad (2)$$

CH2673-2/89/0000-0108 $1.00 © 1989 IEEE

where a_i is the LPC centroid of the frames in segment i. This minimization can be performed employing standard dynamic programming techniques[13]. It should be noted that the segmentation algorithm can also employ distortion measures other than likelihood ratio. In these cases the segmentation will not, however, be optimal in the maximum likelihood sense.

For a given speech utterance, we do not know in advance the number of segments in which to segment. In order to obtain segments of acoustically similar frames, we adopt the algorithm suggested in [9]. Here, the number of segments is steadily increased until the average spectral distortion is smaller than some pre-defined threshold, ε. For the likelihood ratio distortion measure, a reasonable value is $\varepsilon = 0.08$. For cepstral parameters, Euclidian distortion measure we have used $\varepsilon = 0.065$.

3. Segment quantization

The segmentation procedure described in Section 2 produces a large number of acoustic segments which span the speech segment space. Our aim here is to divide this space into N partitions, e.g.; N clusters, and represent each partition by its corresponding cluster center. Here, N is the pre-defined number of sub-word units used in our system. The partitioning is performed by the segment quantization (SQ) algorithm, where the corresponding cluster centroids are computed and a SQ codebook is generated.

The SQ algorithm is based on the assumption that the intra-segment spectral variation is so small that each segment is spectrally well defined by its centroid. The validity of this assumption is controlled by a proper choice of the threshold ε in the automatic segmentation algorithm. The SQ algorithm is briefly described as follows: Given a set of I training segments, $S=\{S_1, S_2, \ldots, S_I\}$ and their corresponding centroids. Let us denote the set of these segment centroids by $C=\{c_1, c_2, \ldots, c_I\}$. Our aim here is to design a codebook of N codevectors, $V=\{v_1, v_2, \ldots, v_N\}$, such that the total distortion

$$D = \sum_{i=1}^{I} \min_{j \in 1,\ldots,N} d(c_i, v_j) \qquad (3)$$

is minimized. Here, $d(c_i, v_j)$ is the distortion between the centroid vector c_i and the codebook vector v_j. The problem of finding the codebook that minimizes (3) is identical to the VQ design problem and can thus be solved by the standard LBG algorithm[12].

The SQ algorithm is applied on the acoustic segments obtained from the data in the training set and the SQ codebooks are designed for pre-defined values of N.

Using the SQ codebook the acoustic segments are then labelled as one of the N sub-words using minimum distortion classification. This gives N segment clusters, one for each codeword.

4. Sub-word modelling.

The acoustic segments labelled as belonging to one of the N sub-word clusters are assumed to be generated by a 1st order Markov model. The topology of the hidden Markov model used in our system is shown in Fig. 1.

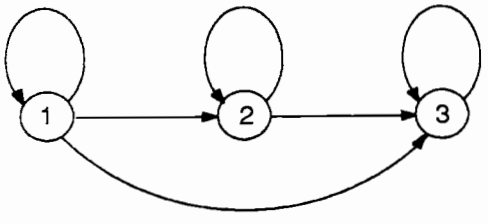

Figure 1. 3-state left-to-right HMM.

The model consists of 3 states and is confined to only left-to-right transitions. The observation probability density function for each state is assumed here to be multivariate Gaussian. For each sub-word cluster, the parameters of the hidden Markov model describing the cluster are computed using the standard Viterbi algorithm[5].

5. Acoustic lexicon generation

Using acoustically defined sub-word units, we no longer have the simple mapping from sub-words to whole words as we would have using phonetically based sub-words combined with a standard phonetic transcription (found in a dictionary). Instead, we have to generate a new "lexicon" where the different words are described in terms of the new acoustic sub-word units. However, the increased accuracy by which segmentation can be performed when an acoustic criterion is used probably outweights this extra complexity.

There are several possible ways of describing an acoustic lexicon for a specified vocabulary. Adhering to the statistical point of view which is followed when HMM modelling is used, a statistical (Markov) network should be an interesting option. Such networks describes each word by a network of sub-word nodes connected by branches which are assigned the probability of using this particular branch.

For our initial work we, however, chose to investigate three simple approaches all of which represent the word references as a base-line sequence of sub-words. These baseline sequences were generated in the following ways:

1) **Random selection.** I references were created by arbitrarily selecting I training utterances which were segmented and labelled. The label sequences then define the sub-word sequences.

2) **Clustering.** For each word in the vocabulary, the training utterances were input to a modified k-means algorithm. The k-means algorithm was based on whole-word DTW and produces I reference cluster centers. The cluster centers obtained for each word were then segment quantized, yielding as a reference a sequence of sub-word template indices. In the recognition phase, the sub-word template indices were used to build sub-word based models of the words in the vocabulary by concatenating the corresponding sub-word models.

3) **Quantized clustering.** Representing each cluster by the segment quantized version of the cluster center introduces a mismatch ("quantization error") between the reference created by the clustering procedure and the actual reference used in the recognition phase. In the quantized clustering procedure we cluster the *segment quantized* versions of the training utterances using the modified k-means algorithm. In order to apply whole-word DTW to the quantized utterances, the segment quantization is performed in the following manner: When creating the SQ codebook, the codebook entries are selected as *actual segment centroids* thereby establishing a corresponding segment codebook. Segmentation and labelling of the training utterances is then performed as previously described. Each labelled segment is then replaced by a linearly warped version of the corresponding segment codebook entry.

6. Recognition

Figure 2 shows the block diagram of the recognition process. Here, the speech parameters (e.g.; LPC, cepstrum) are computed for each frame of the input speech utterance. The parameterized speech utterance is compared with the models of all the words in the vocabulary using the Viterbi decoding algorithm and the recognition is done by applying the maximum likelihood decision rule.

In order to generate the word models, the sequence of sub-words for that word is taken from the acoustic lexicon. The word models are generated by concatenating the corresponding acoustic sub-word HMMs.

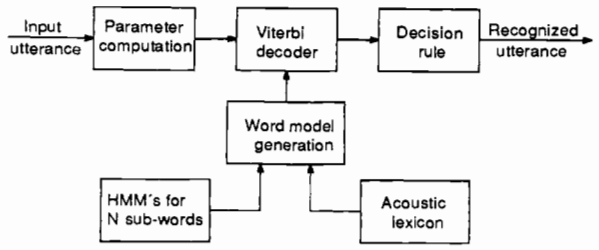

Figure 2. Recognition system

7. Experimental results

7.1 Database

The database consists of 100 repetitions of each of 42 Norwegian words, spoken by one male speaker. The database was divided into two sets of 50 utterances used for training and test respectively. The vocabulary is composed of the 29 letters in the Norwegian alphabet, 10 digits (0-9) and 3 control words ("start", "stopp", "gjenta"). Most of these words are monosyllabic and some are very easily confused also by human listeners. This makes high recognition scores hard to obtain.

The recordings of the database have been performed during a period of 5 weeks. The input speech was sampled at 8 kHz after lowpass filtering with a cut-off frequency of 3.5 kHz. The recordings were done in an ordinary office with a low level of background noise.

A block diagram of the preprocessing chain is shown in Fig. 3. First, the speech is filtered by an IRS filter to mimic the average telephone microphone characteristic. The signal is then pre-emphasized with the filter $H(z) = 1 - 0.95\ z^{-1}$. Estimates of 11 autocorrelation coefficients were then obtained every 15 ms using overlapped frames of 45 ms, weighting each frame by a Hamming window.

Each input utterance was recorded using a 2 second recording interval. In order to extract the speech information, endpoint detection is needed. The algorithm employed is energy based and is described in [15]. Because of different recording conditions, the thresholds were reoptimized. Some problems were encountered with the words "start" and "stopp". Probably due to the background noise, the initial /s/ was not detected properly. Since the main purpose of this study was to investigate the performance of the recognition system, we chose to use special thresholds for these words. This was in order to be able to assume that the endpoints were correctly detected and that the endpoint detection algorithm had no influence on system performance.

7.2 Experiments

In order to obtain a reasonable estimate of the system's performance with a minimum of computational load, the system has been exhaustively tested on a small 3-word vocabulary consisting of the Norwegian letters /b, d, g/. This vocabulary is selected because it is reasonably complex as it has high confusability, and it is small enough to allow for fast computation of results.The number of sub-word templates in the codebook is N=16 and one reference/word was used. The HMM models in all cases use a diagonal covariance matrix.

In our first experiment we compared two of the methods for acoustic lexicon generation, random selection and clustering. After preprocessing, 10 LPC coefficients were extracted and the Likelihood ratio distortion measure was used for segmentation, labelling and acoustic lexicon generation. The LPC coefficients were converted to 10 cepstral coefficients which were used for HMM model building. Testing the two methods, clustering proved to give a considerable performance improvement over random selection, 84.7% and 72.7% correct recognition respectively.

Next, we wished to investigate the effects of distortion measure on the performance of the system. In this experiment the clustering method for acoustic lexicon generation was employed. The distortion measures investigated were Likelihood ratio (LR) using 10 LPC coefficients, Euclidian distortion measure using 14 LPC-cepstrum parameters and Euclidian distortion measure using 14 liftered cepstral coefficients[16]. For the HMM model building, cepstral coefficients were used, 10 coefficients for LR, 14 coefficients otherwise. The results are shown in Table 1.

Distortion measure	Recognition rate
Likelihood ratio	84.7%
Cepstrum	77.3%
Liftered cepstrum	90.0%

Table 1. Recognition results, clustering.

As can be seen from the table, the distortion measure has considerable effect on the performance. The use of liftering, which has the effect of smoothing the power spectrum clearly improves the performance of the sub-word based recognition system.

We also wished to perform a comparison between the clustering and the quantized clustering method for acoustic lexicon generation. Using the same distortion measures as described above, the experiment was repeated for the quantized clustering method. The results are shown in Table 2.

Distortion measure	Recognition rate
Likelihood ratio	96.7%
Cepstrum	84.7%
Liftered cepstrum	94.0%

Table 2. Recognition results, quantized clustering.

The use of quantized clustering improved the system performance for all distortion measures employed. Surprisingly, the use of likelihood ratio now yielded better results than the use of liftered cepstrum. This might be due to the fact that employing the quantized training utterances in the clustering procedure introduces a smoothing effect in the training phase.

In the sub-word based speech recognition systems described so far, we have represented acoustic segments by their centroids both in the maximum likelihood segmentation procedure and in the segment quantization procedure. Representation of acoustic segments by their centroids captures only the static information about the segment. Here,

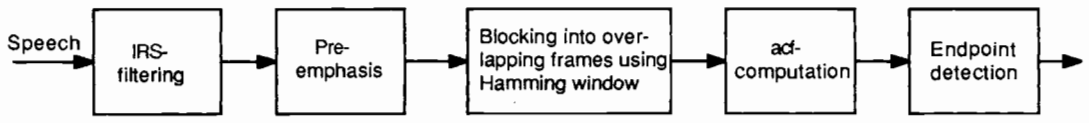

Figure 3. Preprocessing

we suggest that using dynamic information in addition to static information in the representation of acoustic segments, we can improve the recognition performance of the acoustic sub-word based speech recognition system further.

In order to test this hypothesis, we use the cepstrum coefficients and the delta-cepstrum coefficients[14] derived through LPC analysis for parameterization. Use of the delta-cepstrum coefficients provides information about dynamic spectral features of the segments. Using the parameterization in terms of cepstrum and delta-cepstrum coefficients, we have again performed segmentation, segment quantization and HMM model building for the sub-words. A diagonal covariance matrix is used in the HMM model building procedure. The acoustic lexicon is generated using the modified k-means algorithm. The recognition score is found to be 98.0% Using liftered cepstral parameters and delta-cepstrum, the recognition rate was 94.7%. We see that the inclusion of dynamic spectral information to represent the acoustic segments improves the performance of the system significantly.

In order to compare the results of our sub-word based system, we have implemented baseline whole-word based systems using both the DTW and HMM approach. The recognition scores for these systems were 97.3% and 90.0% respectively. Thus, the best sub-word based recognizers perform as well as the best base-line whole-word recognizers.

Finally, the systems were evaluated for the full vocabulary. The results are summarized in Table 3.

System	Recognition rate
Whole-word HMM	91.0%
Whole-word DTW	97.0%
ASWU	90.8%

Table 3. Results for 42-word vocabulary.

The results for the full vocabulary confirm the findings for the /b,d,g/ vocabulary. The sub-word based system performs as well as the HMM-based whole-word recognizer but is inferior to the DTW-based whole word recognizer. A reason for the DTW-based recognizer's high performance may be the extreme consistency of speaking for the one speaker in the database. Statistics of the warping path show that an overwhelming majority of the utterances are spoken with small deviations in temporal structure.
The inclusion of transitional spectral information has not yet been investigated for the full vocabulary. It is expected that this will boost the performance of the sub-word unit based system.

8. Summary

We have presented a sub-word based speech recognition system. The system has been compared to baseline whole-word based HMM and DTW speech recognizers. Experiments have shown that our system performs as well as the HMM-based recognizer but is inferior to the DTW-based system. However, our database consisted of utterances spoken by only one speaker and further experiments must be undertaken before definitive conclusions can be drawn.

Although this system was evaluated for a 42-word vocabulary, it might be noted that the acoustic sub-word unit based approach is specially well suited for large vocabulary speech recognition where it not only requires a smaller amount of training data, but also reduces the computational complexity with respect to the DTW-based whole-word based system.

Acknowledgements.

This work has been performed under ELAB's multiclient speech processing project financed by the Norwegian Telecommunications Administration, The Royal Norwegian Council for Scientific and Industrial Research, EB Technology A/S, STK-ALCATEL A/S and STENTOFON A/S.

References

[1] L.R.Rabiner, S.E.Levinson: "Isolated and connected word recognition - Theory and selected applications", IEEE Trans. COM-29, pp.621-629, May 1981

[2] J.G.Wilpon, D.M.DeMarco,R.P.Mikkilineni: "Isolated word recognition over the DDD telephone network - Results of two extensive field studies", Proc. ICASSP, pp. 55-58, April 1988

[3] A.Averbuch et al.: "Experiments with TANGORA 20,000 word speech recognizer", Proc. ICASSP, pp. 701-704, April 1987

[4] L.R.Rabiner, J.G.Wilpon, F.K.Soong: "High performance connected digit recognition using hidden Markov models", Proc. ICASSP, pp. 119-122, April 1988

[5] L.R.Rabiner, B.H.Juang: "An introduction to hidden Markov models", IEEE ASSP Magazine, pp. 4-16, Jan. 1986

[6] L.R.Bahl et al.: "Recognition results with several experimental acoustic processors", Proc. ICASSP, pp. 249-251, April 1979

[7] K.K.Paliwal, A.M.Kulkarni: "Segmentation and labelling using vector quantization and its application in isolated word recognition", Journ. Acoust. Soc., India, Vol. 15, pp. 102-110, Jan. 1987

[8] J.G.Wilpon, B.H.Juang, L.R.Rabiner: "An investigation on the use of acoustic sub-word units for automatic speech recognition", Proc. ICASSP, pp.821-824, April 1987

[9] C.H.Lee, F.K.Soong, B.H.Juang: "A segment model based approach to speech recognition", Proc. ICASSP, pp. 501-504, April 1988

[10] L.R.Bahl et al.: "Acoustic Markov models used in the TANGORA speech recognition system", Proc. ICASSP, pp. 497-500, April 1988

[11] V.R.Algazi, K.L.Brown: "Automatic speech recognition using acoustic sub-words and no time alignment",

[12] Y.Linde, A. Buzo, R.M.Gray: "An algorithm for vector quantizer design", IEEE Trans. COM-28, pp. 84-95, Jan. 1980

[13] T.Svendsen, F.K.Soong: "On the automatic segmentation of speech signals", Proc. ICASSP, pp. 77-80, April 1987

[14] F.K.Soong, A.E.Rosenberg: "On the use of instantaneous and transitional spectral information in speaker recognition", Proc. ICASSP, pp. 877-880, April 1986

[15] E.Harborg, M.H.Johnsen, T.Svendsen: "Speaker independent recognition of isolated Norwegian words" (in Norwegian), ELAB Report STF44F86064, June 1986

[16] B.H.Juang, L.R.Rabiner, J.G.Wilpon:"On the use of bandpass liftering in speech recognition", IEEE Trans. ASSP-35, pp. 947-954, July 1987.