INCORPORATING LINGUISTIC KNOWLEDGE AND AUTOMATIC BASEFORM GENERATION IN ACOUSTIC SUBWORD UNIT BASED SPEECH RECOGNITION

Trym Holter and Torbjørn Svendsen

Department of Telecommunications, Norwegian University of Science and Technology
O.S. Bragstads plass 2B, N-7034 Trondheim, Norway
E-mail: holter@tele.ntnu.no or svendsen@tele.ntnu.no
Tel.:+47 73 594318. Fax: +47 73 592640.

10.21437/Eurospeech.1997-13

ABSTRACT

A major challenge in speech recognition based on acoustic subword units is creating a lexicon which is robust to inter- and intra-speaker variations. In this paper we present two different approaches for incorporating simple word-level linguistic knowledge into the labelling step of the training procedure. The proposed systems also utilise a scheme for combined optimisation of baseforms and subword models. For the TI46 database, these methods are shown to greatly improve the performance compared to an acoustic subword based speech recogniser employing unsupervised labelling, and they are found to perform as well as systems utilising whole-word models and context independent phoneme models.

1. INTRODUCTION

Traditionally, automatic speech recognisers employ phone-like units based upon a linguistic description of the language. On the other hand, the analysis of the actual speech signal is acoustically based. The resulting system is neither phonetically nor acoustically consistent, but is instead a hybrid of two methodologies. In an attempt to create a consistent acoustic framework, there have been several attempts to utilise acoustically based subword units (ASWUs) over the last ten years (see e.g., [1, 2, 3]).

One major challenge with this kind of basic units is the lack of a pronunciation lexicon. The lexicon should contain one or several baseforms for each word in the vocabulary. Each baseform defines the composition of a word in terms of the basic units. As the ASWUs do not necessarily have a one-to-one correspondence to any linguistic units, the baseforms must be found by some training procedure, e.g., such as proposed in [4, 5].

This work concentrates on speaker independent recognition. Most algorithms for ASWU based speech recognisers are developed and tested for speaker dependent recognition, which reduces the problem of inter-speaker variations. Since the ASWUs are automatically defined, based on the acoustic manifestations in the training material, an acoustic segmentation and a subsequent labelling of the resulting segments are required. The labelling will generally be sensitive to inter- and intra-speaker variations, and has been the weak link when applying ASWUs to speaker independent recognition. In this paper, we propose two different approaches for labelling of acoustically segmented speech. The two procedures incorporate simple word-level linguistic knowledge in the training phase by restricting the labelling of different utterances of the same word to be similar or identical.

2. BASIC TRAINING SCHEME

The basic training scheme closely resembles that proposed in [2]. The major differences lie in the labelling schemes, and in how the baseforms are generated. The training of the proposed system can be summarised as follows:

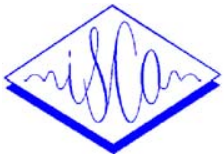
1) Initial segmentation of speech utterances into acoustically stationary segments is performed by Constrained Clustering Segmentation [6]. Choosing Euclidean distance as the distortion measure, this problem can be formulated as finding the set of segment boundaries $\{b_0, b_1, \ldots, b_J\}$ that minimises the total distortion

$$\sum_{j=0}^{J-1} \sum_{n=b_j}^{b_{j+1}-1} \|\mathbf{x}_n - \bar{\mathbf{x}}_j\|^2, \qquad (1)$$

where $\bar{\mathbf{x}}_j$ is the centroid of the j'th segment consisting of the feature vectors $\{\mathbf{x}_{b_j}, \ldots, \mathbf{x}_{b_{j+1}-1}\}$.

2) Representing each of the acoustic segments by its centroid, the LBG-algorithm [7] is employed to cluster the segments into S clusters, and to create a corresponding codebook of S codewords. S is the predefined number of subword units used in our system.

3) Labelling of the acoustic segments into subword classes on basis of the codebook from step 2 is done by one of the two labelling procedures. The purpose is to incorporate some word-level linguistic knowledge, and thereby increase the robustness to inter-



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

and intra-speaker variations. The methods are further described in section 3.

4) Each partition of the feature space resulting from the clustering, represents an ASWU. For each subword a Hidden Markov Model (HMM) is trained from the acoustic segments in the corresponding cluster.

5) The baseforms, in terms of the ASWUs, are generated. The baseform optimisation method is based on a Maximum Likelihood (ML) formulation [5] and relies on the Modified Tree-Trellis algorithm [4]. This step should include a combined optimisation of the HMMs and the baseforms as described in [8].

3. LABELLING OF ACOUSTICALLY SEGMENTED SPEECH

In previous work, the labelling in terms of subword units has been completely unsupervised, based upon the results of the clustering performed in step 2 of the training procedure. In this case, the label assigned to segment j corresponds to the label of the codeword which minimises the distance over the frames in the segment, $\{\mathbf{x}_{b_j}, \ldots, \mathbf{x}_{b_{j+1}-1}\}$. Using the minimum squared error criterion, the optimal label i' is found according to:

$$i' = \underset{i \in \mathcal{I}}{\arg\min} \sum_{n=b_j}^{b_{j+1}-1} \|\mathbf{x}_n - \mathbf{c}_i\|^2, \quad \mathbf{c}_i \in C, \qquad (2)$$

where C is the codebook, $\mathbf{c}_i$ is the codebook vector corresponding to index i, and $\mathcal{I}$ is the set of indices in the codebook, $\mathcal{I} = \{0, 1, \ldots, S-1\}$.

Inter- and intra-speaker acoustic variability may cause this procedure to yield very different label strings for different utterances of the same word. Because of this, a single representative lexical description of each word is difficult to find, and investigations have shown that the ML based baseform optimisation (step 5) does not perform well under these circumstances.

In sections 3.1 and 3.2, we propose two different labelling procedures that utilise knowledge of which words that constitute a training utterance.

3.1. Centroid-alignment-constrained labelling

This scheme aims to assign *similar* label-sequences to all utterances found of the same word in the training data. The algorithm can be divided into a centroid-alignment procedure (step 1–3 below) followed by a constrained labelling (step 4 below). The centroid-alignment procedure ties segments from different utterances of the same word to each other by use of Dynamic-Time-Warping (DTW). Each utterance is represented as a sequence of centroids, one per acoustic segment. The DTW algorithm performs alignment of two sequences by searching for the optimal path through a grid of nodes. Each node represents a pair of centroid vectors. The cost associated with a node (p_l, q_l) is given by

$$d_N(p_l, q_l) = \|\bar{\mathbf{x}}_{p_l}^{(r)} - \bar{\mathbf{x}}_{q_l}^{(t)}\|^2, \qquad (3)$$

where $\bar{\mathbf{x}}_{p_l}^{(r)}$ and $\bar{\mathbf{x}}_{q_l}^{(t)}$ are the centroids of segment p_l in utterance r and q_l in utterance t, respectively. The entire match is found by

$$D = \frac{\sum_{l=1}^{L} d_N(p_l, q_l) \cdot d_T(p_l, q_l | p_{l-1}, q_{l-1})}{f(r,t)}, \qquad (4)$$

where $d_T(p_l, q_l | p_{l-1}, q_{l-1})$ is the cost associated with the transition from node (p_{l-1}, q_{l-1}) to node (p_l, q_l), and $f(r,t)$ is a distance normalisation factor. There exists a variety of strategies for choosing d_T and f. In the present work, d_T is chosen as

$$d_T(p, q | u, v) = \begin{cases} 1, & u = p-1 \wedge v = q-1 \\ \alpha, & u = p-2 \wedge v = q-1 \\ \beta, & u = p-1 \wedge v = q-2 \\ \infty, & \text{otherwise}, \end{cases} \qquad (5)$$

where $\alpha, \beta \geq 1$ are constants that decide the penalty associated with deletions and insertions, respectively. The normalisation factor $f(r,t)$ is chosen to equal the number of segments in the utterance $x^{(t)}$.

The DTW algorithm is the core of the proposed labelling scheme described below:

For each word:

1) Represent each training token as a sequence of centroid vectors

2) Find a reference token as the utterance that has the smallest average DTW-distance to all other utterances of that word.

3) Employ DTW to align each training token with the reference token. All acoustic segments aligned with the same segment of the reference constitute a cluster. Acoustic segments not aligned with any reference segment make up a single-member cluster.

4) Label each cluster by the index of the nearest codeword according to

$$i' = \underset{i \in \mathcal{I}}{\arg\min} \sum_{m=1}^{M} \|\bar{\mathbf{x}}_{p_m}^{(k_m)} - \mathbf{c}_i\|^2, \quad \mathbf{c}_i \in C, \qquad (6)$$

where M is the size of the cluster. Each cluster member is identified by the segment number (p_m) and utterance number (k_m).

3.2. Joint resegmentation and labelling

The idea of this scheme is to assign *identical* label-sequences to all utterances found of the same word in the training data. The label sequence *and* the segment boundaries in each utterance should be chosen so that an overall objective criterion is minimised. Utilising the minimum squared error criterion, the optimal number of segments J' and the optimal label sequence $\{i'_0, \ldots, i'_{J'-1}\}$ is given by

$$\{i'_0, \ldots, i'_{J'-1}\} =$$

$$\arg\min_{\{i_j \in \mathcal{I}, J \in \mathbb{N}\}} \sum_{k=0}^{K-1} \min_{\{b_j^{(k)}\}} \sum_{j=0}^{J-1} \sum_{n=b_j^{(k)}}^{b_{j+1}^{(k)}-1} \|\mathbf{x}_n^{(k)} - \mathbf{c}_{i_j}\|^2, \quad (7)$$

$$\mathbf{c}_{i_j} \in C,$$

where

K	is the number of utterances,
$\mathbf{x}_n^{(k)}$	is frame n in utterance k,
$b_j^{(k)}$	is segment boundary j in utterance k, and
C	is the codebook created from the initial acoustic segmentation.

This optimisation problem can be expressed as a search through a trellis, where each state in the trellis is associated with one codeword in the codebook. The path through the trellis should be chosen so that the overall distance with regard to all utterances is minimised according to equation 7. In this framework, additional requirements regarding the minimum duration of each subword unit is easily incorporated.

The solution to the given problem can be found by the Modified Tree-Trellis algorithm [4]. The problem formulation is very similar to the joint log-likelihood maximisation traditionally performed by the Modified Tree-Trellis algorithm. There are two main differences. First, the probability density function calculation traditionally associated with each state is replaced by an Euclidean distance calculation. Second, no transition probabilities are included in the distance score. The transitions only describe the legal successors of each node in a state-space description.

In the optimisation described by equation 7, the optimal segment boundaries in each utterance are found as a by-product. These boundaries are needed for training of the HMMs, as described by step 4 of the training procedure.

Due to the large search-space, the Modified Tree-Trellis algorithm may suffer from memory shortage. In the present work we have utilised the Extended Multiple Candidate Method [5] to constrain the search-space. This is a two-step procedure. First the search space is constrained by a ML-search for the N best baseforms for each utterance of the given word. Now, the solution to the baseform selection problem is the one baseform that maximises the joint likelihood of all utterances of the word. The size of N in the N-best search will affect the quality of the resulting lexicon. In [5], experiments on a phonemic subword based speech recogniser showed little increase in performance for $N > 5$.

4. EXPERIMENTS

The proposed systems have been tested on the database TI20, a subset of TI46. This corpus contains 16 speakers. The vocabulary consists of 20 words; the digits and ten computer-related words. It is chosen for these initial experiments because it contains a small number of words uttered in isolation, which eliminates the problem of identifying the word boundaries for use in the training phase. The database also contains enough different speakers to give a reasonable inter-speaker acoustic variability.

For the initial segmentation, a distortion threshold is required. As proposed in [2], this was set to $\epsilon = 0.065$ for a system utilising 14 linear prediction derived cepstral coefficients. This feature set was used for segmentation and labelling. For modelling and recognition we chose to use a more standard 39 component feature vector as in [8]. Feature vectors were extracted every 15 ms using a 45 ms Hamming window in both cases. In the segmentation procedures, the duration of a subword was constrained to be at least 30 ms.

The Extended Multiple Candidate Method with $N = 20$ was used both in the joint resegmentation and labelling scheme and for baseform optimisation with both labelling schemes.

The ASWUs were modeled by one-state HMMs. Experiments were performed with one to five components in the Gaussian mixture pdfs. For each model set, two iterations of the combined optimisation of baseforms and subword models were performed, resulting in a different number of distinct subword units in the different lexicons. The upper limit of subword units is S, which is the size of the codebook. For small vocabularies, the baseforms will often be composed of less than S different ASWUs in total. In these experiments, S was set to 128, and the resulting number of units after baseform optimisation ranged from 102 to 110.

In addition to the proposed ASWU-based systems, three reference systems were designed. These systems were based on whole-word models, phoneme subword models, and acoustic subword units, respectively. The ASWU based system employed an unsupervised labelling scheme, and no combined optimisation of baseforms and subword units was performed. This system is, except for the baseform optimisation method, very similar to the systems proposed in [1, 2]. All systems were tested with different numbers of components in the Gaussian mixture pdfs. An overview of the tested systems is given below:

(a) The proposed ASWU-based system employing the centroid-alignment-constrained labelling scheme. 102-110 models were utilised with 1 state per unit and 1-5 mixtures per state. Two iterations of combined optimisation of HMMs and baseforms were performed.

(b) The proposed ASWU-based system employing the joint resegmentation and labelling scheme. 103-105 models were utilised with 1 state per unit and 1-5 mixtures per state. Two iterations of combined opti-

misation of HMMs and baseforms were performed.

(c) A system based on whole-word models. 20 models were used, each with 7 states per word and 1-3 mixtures per state.

(d) A system based on context independent phoneme models. 29 models were used, each with 3 states per phoneme and 1-5 mixtures per state. A standard lexicon was utilised.

(e) A system based on ASWUs with 1 state per unit and 1-5 mixtures per state. The training of this system followed the scheme outlined in section 1, except for step 3. The labelling was performed in an unsupervised manner, as described by equation 2. No combined optimisation of HMMs and baseforms was performed.

In table 1 word recognition rates and the number of free parameters in the HMMs are shown for the five different systems.

Mix.	(a)	(b)	(c)	(d)	(e)
1	98.0%	98.7%	93.9%	98.0 %	75.0%
	8690	8216	11060	6873	8137
2	98.4%	99.5%	99.4%	99.3 %	69.4%
	16590	16274	22120	13746	16116
3	99.3%	99.9%	99.7%	99.6 %	73.3%
	24648	24648	33180	20619	24411
4	99.7%	99.6%	*not*	99.6 %	70.8%
	32232	32864	*tested*	27492	33496
5	99.7%	99.9%	*not*	99.8 %	72.5%
	41870	41475	*tested*	34365	41080

Table 1: Word correct rate and number of free parameters for the described systems.

The results in table 1 show that the constrained labelling procedures of section 3 together with the combined optimisation scheme for baseform generation and subword modelling clearly improve the performance compared to an ASWU based system which does not include these methods. This is mainly due to the labelling schemes, which ensure that all utterances of a given word are labelled similarly or identically. For speaker independent systems, we believe an unsupervised labelling scheme will result in a too inconsistent labelling, which is not a good basis for the model estimation.

The proposed ASWU based systems should also be compared to the systems based on whole-word models and phoneme subword models. The performance of these systems relative to each other is hard to analyse on this task, as they all give recognition rates very close to 100%. Even though the systems must be said to perform equally well for this corpus, the proposed systems should be tested on a speaker independent database of larger complexity to give certain conclusions. This is subject to further research.

5. CONCLUSIONS

We have presented two different methods for incorporating simple word-level linguistic knowledge into the labelling of acoustically segmented speech. A combined optimisation scheme for baseform generation and subword modelling has also been investigated in the context of speech recognition utilising ASWUs. These procedures resulted in greatly improved system performance. The proposed systems gave recognition rates comparable to systems based on whole-word modelling and phonemic subword modelling.

6. REFERENCES

[1] C.-H. Lee, B.-H. Juang, F. K. Soong, and L. Rabiner, "Word recognition using whole word and subword models," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, (Glasgow, Scotland), pp. 683–686, IEEE, May 1989.

[2] T. Svendsen, K. Paliwal, E. Harborg, and P. O. Husøy, "An improved sub-word based speech recognizer," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, (Glasgow, Scotland), pp. 108–111, IEEE, May 1989.

[3] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer, and M. A. Picheny, "A method for the construction of acoustic Markov models for words," *IEEE Trans. Speech and Audio Processing*, vol. 1, pp. 442–452, Oct. 1993.

[4] T. Svendsen, F. K. Soong, and H. Purnhagen, "Optimizing baseforms for HMM-based speech recognition," in *Proc. European Conf. on Speech Commun. and Techn. (EUROSPEECH)*, (Madrid, Spain), pp. 783–786, Sept. 1995.

[5] T. Holter and T. Svendsen, "A comparison of lexicon-building methods for subword-based speech recognisers," in *Proc. IEEE Region 10 Conf. on Digital Signal Proc. (TENCON)*, (Perth, Australia), pp. 102–106, IEEE, Nov. 1996.

[6] T. Svendsen and F. K. Soong, "On the automatic segmentation of speech signals," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, (Dallas, USA), pp. 77–80, IEEE, Apr. 1987.

[7] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, pp. 84–95, Jan. 1980.

[8] T. Holter and T. Svendsen, "Combined optimisation of baseforms and subword models for an HMM based speech recogniser," in *Proc. The 4th Int. Symposium on Signal Processing and its Applications (ISSPA)*, (Gold Coast, Australia), pp. 321–324, Aug. 1996.