

A Joint Segmentation and Labelling Scheme for use in Acoustic Subword 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.

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 a joint segmentation and labelling scheme to incorporate word-level linguistic knowledge into the training procedure. The proposed system is also based on a combined optimisation of the baseforms and the subword models. For the TI46 database, this method is shown to greatly improve the performance compared to an acoustic subword based speech recogniser employing unsupervised labelling, and is found to perform as well as systems utilising whole-word models and context independent phoneme models.

1. INTRODUCTION

Traditionally, the subword units employed in automatic speech recognition have been defined based upon a linguistic description of the language. A major disadvantage of this approach is the inherent mismatch between the acoustically based analysis of the actual speech signal and the linguistically based description of the language. Over the last ten years, there have been attempts to utilise acoustically based subword units (ASWUs) in automatic speech recognition to avoid this mismatch (see e.g., [1, 2, 3]).

One major challenge with this kind of basic units is the lack of an utterance lexicon. The lexicon should contain one or several baseforms for each word in the vocabulary. Each baseform define the composition of a word in terms of the basic units. For linguistically based subword units, the lexicon is most often extracted from a standard pronunciation dictionary or created by an human expert. As the ASWUs do not necessarily have a one-to-one correspondence to any linguistic units, these procedures are not feasible for ASWU-based systems. This means that the baseforms must be found by some training procedure, such as e.g. 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. These methods do not perform well for speaker independent recognition. In this paper, we propose a method to incorporate word-level linguistic knowledge in the training phase through a joint segmentation and labelling scheme. This is intended to reduce the problems of inter- and intra-speaker variations, and to improve the automatic baseform generation.

The paper is organised as follows. In section 2, the basic training scheme is briefly described. In section 3 the joint segmentation and labelling scheme is discussed. In section 4 results from tests on the TI46 database [6] are reported and finally some conclusions are given in section 5.

2. BASIC TRAINING SCHEME

The basic training scheme closely resembles that of [2]. The major differences lie in the joint re-segmentation and labelling scheme 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 by the use of Constrained Clustering Segmentation [7]. Choosing Euclidean distance as the distortion measure, this problem can be formulated as finding the segment boundaries $\{b_0, b_1, \dots, b_J\}$ that minimises the total distortion

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

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

2. Clustering of the acoustic segments into M clusters, where M is the predefined number of subword units used in our system. For this, each segment is represented by its centroid. A codebook with M codewords is created on basis of the clusters. The LBG-algorithm [8]

is utilised for the clustering and codebook design.

3. Joint resegmentation and labelling of the acoustic segments into subword classes on basis of the codebook from step 2. The purpose is to incorporate some word-level linguistic knowledge, and thereby increase the robustness to inter- and intra-speaker variations. The method is 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. Initial models are estimated by the Segmental K-means procedure. Further training is performed utilising the Baum-Welch re-estimation scheme.
5. Generation of baseforms, in terms of the ASWUs. 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 [9].

3. JOINT RESEGMENTATION AND LABELLING OF 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 codeword which minimises the distance over the frames in the segment, $\{\mathbf{x}_{b_j}, \dots, \mathbf{x}_{b_{j+1}-1}\}$. Given the minimum squared error criterion, the optimal label i' is found according to:

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

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, $\mathcal{I} = \{0, 1, \dots, M-1\}$.

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 the present work, knowledge of which word is uttered is incorporated in a joint resegmentation and labelling procedure. The idea is to assign the same sequence of labels 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 $\{i'_0, \dots, i'_{J'-1}\}$ is given by

$$\{i'_0, \dots, i'_{J'-1}\} = \operatorname{argmin}_{\{i_j \in \mathcal{I}, J \in \mathbb{N}\}} \sum_{k=0}^{K-1} \min_{\{b_j^{(k)}\}} \sum_{j=0}^{J-1} \sum_{n=b_j^{(k)}}^{b_{j+1}^{(k)}-1} \|\mathbf{x}_n^{(k)} - \mathbf{c}_{i_j}\|^2, \quad (3)$$

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

where

- K is the number of utterances for the given word,
- $\mathbf{x}_n^{(k)}$ is frame n in utterance k , and
- $b_j^{(k)}$ is segment boundary j in utterance k .

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 3. In this framework, additional requirements regarding the minimum duration of each subword unit is easily incorporated. The state-diagram in figure 1 shows a configuration where the minimum allowed duration of each subword unit is two frames. The symbol “ \emptyset ” denotes a null transition, i.e., a transition that is not associated with an observable event.

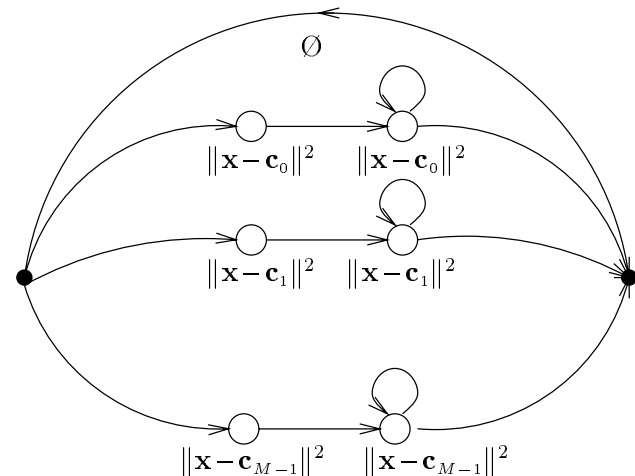


Figure 1. State-diagram describing the possible sequences of subword labels.

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 3, 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 system has 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. However, this is not a completely speaker independent task, as the same speakers are found in both test- and training data. It should therefore be considered a *multiple speaker* task.

Different preprocessing strategies were chosen for the segmentation/labelling steps, and for the modelling/recognition steps. In both cases, the speech was first pre-emphasised utilising the filter $H(z) = 1 - 0.95z^{-1}$. For segmentation and labelling, 14 cepstral coefficients were extracted from the linear predictive coefficients. For modelling and recognition, feature vectors containing 39 parameters (12 mel-frequency cepstral coefficients and normalised log energy, plus first and second order difference coefficients) were computed from each frame. Feature vectors were extracted every 15 ms using a 45 ms Hamming window in both cases.

In the labelling procedure, two frames was chosen as the minimum duration of a subword unit. For the initial segmentation into acoustically stationary segments, a distortion threshold $\epsilon = 0.065$ [2] was set. In the joint resegmentation and labelling step, no such threshold is necessary. The final number of segments will nevertheless be related to the chosen threshold. In these experiments, the number of segments increased by 19% compared to the initial segmentation after the joint resegmentation and labelling. In both the joint resegmentation and labelling procedure and in the baseform optimisation step, the Extended Multiple Candidate Method was used with $N = 20$ in the initial search for baseform candidates.

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 subword units in the different lexicons. The upper limit of subword units is M , which is the size of the codebook. For small vocabularies, the baseforms will often be composed of less than M different ASWUs in total. In these experiments, M was set to 128, and the resulting number of units after baseform optimisation ranged from 103 to 105.

In addition to the proposed ASWU-based system, three reference systems were designed. These systems were based on whole-word models, phonemic subword models, and acoustic subword units, respectively. The ASWU based system was trained without the joint resegmentation and labelling scheme and without the combined optimisation of baseforms and subword units. This system is, except for the baseform optimisation method, very close 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 different systems is given below:

- (a) The proposed ASWU-based system. 103-105 models were utilised with 1 state per unit and 1-5 mixtures per state. The joint resegmentation and labelling scheme and the scheme for combined optimisation of HMMs and baseforms were used.
- (b) A system based on whole-word models. 20 models were utilised, each with 7 states per word and 1-3 mixtures per state.
- (c) A system based on context independent phoneme models. 29 models were utilised, each with 3 states per phoneme and 1-5 mixtures per state. A standard lexicon was utilised.
- (d) 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 joint resegmentation and labelling was replaced by an unsupervised labelling, as described by equation 2. No combined optimisation of HMMs and baseforms was performed.

The total number of free parameters in a system is determined by the number of subword- or word-models, the dimension of the feature vectors, the number of mixtures, and the number of states in each model. For fairness, systems with approximately the same number of free parameters should be compared. In table 1 word recognition rates and the number of free parameters utilised in the HMMs are therefore shown for the four different systems:

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

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

The results in table 1 show that the joint resegmentation and labelling scheme together with the combined optimisation of baseforms and subword HMMs clearly improves the performance compared to an ASWU based system which does not include these procedures. This is mainly due to the joint resegmentation and labelling, which incorporates word-level linguistic knowledge. This procedure ensures that all utterances of a given word are labelled identically. For speaker independent systems, we believe an unsupervised labelling scheme will result in an inconsistent labelling, which is not a good basis for the model estimation.

The proposed ASWU based system should also be compared to the systems based on whole-word models and phonemic 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 system must 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 a joint labelling and segmentation scheme for use with a speech recogniser

utilising acoustic subword units. Together with a scheme for combined optimisation of baseforms and subword HMMs, this algorithm has been incorporated into a procedure for training of ASWU based recognisers, and showed to give a greatly improved performance. The proposed system gave a recognition performance comparable to systems based on whole-word modelling and phonemic subword modelling.

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 Processing*, (Perth, Australia), pp. 102–106, IEEE, Nov. 1996.
- [6] G. R. Doddington and T. B. Schalk, "Speech recognition: Turning theory to practice," in *IEEE Spectrum*, pp. 26–32, Sept. 1981.
- [7] 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.
- [8] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, pp. 84–95, Jan. 1980.
- [9] 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.