A HYBRID ARCHITECTURE FOR AUTOMATIC SEGMENTATION OF SPEECH WAVEFORMS

Iosif Mporas, Todor Ganchev and Nikos Fakotakis

Artificial Intelligence Group, Wire Communications Laboratory,
Dept. Electrical and Computer Engineering, University of Patras, 26500 Rion-Patras, Greece
{imporas, tganchev, fakotaki}@wcl.ee.upatras.gr

ABSTRACT

In the present work, we propose a hybrid architecture for automatic alignment of speech waveforms and their corresponding phone sequence. The proposed architecture does not exploit any phone boundary information. Our approach combines the efficiency of embedded training techniques and the high performance of isolated-unit training. Evaluating on the established for the task of phone segmentation TIMIT database, we achieved an accuracy of 83.56%, which corresponds to improving the baseline system's accuracy by 6.09 %.

Index Terms— Speech segmentation, hidden Markov models, embedded training, isolated-unit training.

1. INTRODUCTION

The development of large speech corpora, consisting of realistic and unconstrained speech, speeded up the evolution of speech technology significantly [1]. Specifically, in the last two decades a number of speech processing tasks, such as text-to-speech conversion [2], voice transformation [3], text-dependent speaker recognition [4, 5] and language identification based on phoneme recognition [6] took advantage of these corpora. This progress allowed the emergence of commercial applications. All these tasks depend on speech corpora with available annotation of speech label boundaries (most often phones, but also syllables, words, sentences). Moreover, the quality of the label alignment is crucial for enhancing the performance of these applications.

Currently, manual annotation of speech recordings to phonemic labels and boundaries is the most accurate method [7]. However, manual annotation is time-consuming (especially for large databases) and expensive (usually expert phonetician is required). Moreover, manual labeling introduces annotators' subjectivity [8]. For these reasons, automatic phoneme segmentation (APS) algorithms have been proposed.

There are two major categories of methods for APS: implicit and explicit [9]. In implicit (or text-independent) techniques, the speech signal is segmented to fragments, corresponding to phone-like (or syllable-like) units, without any knowledge of the corresponding phonetic transcription. In explicit (or text-dependent) methods the speech signal is segmented and time-aligned against a known phonetic transcription. Generally, explicit techniques achieve higher performance, since the number of detected segments is equal to the given in the transcription, in contrast to the implicit case were the number of predicted boundaries is not always correct.

The superior performance of explicit techniques, established them as a traditional choice in the task of creating phonemic transcriptions of speech databases, for which the word level transcription is usually known. Numerous explicit segmentation approaches have been proposed. In [10] synthetic speech is aligned against natural speech, using the dynamic time warping (DTW) algorithm. In [11] a phoneme alignment algorithm based on discriminative learning is proposed. In [12] automatic alignment is carried out using neural networks followed by boundary refinement using heuristic speech-specific knowledge. Most of the reported work on the task of explicit segmentation is based on the well established and widely used hidden Markov models (HMMs) [8, 13, 14].

A typical structure of the HMM-based APS is illustrated in Fig. 1. In HMM-based APS a phoneme recognizer is employed for segmenting the data. Specifically, the text transcription of the speech utterance is converted to the corresponding phone sequence, using a letter-to-sound converter. Subsequently, the speech waveforms are force-

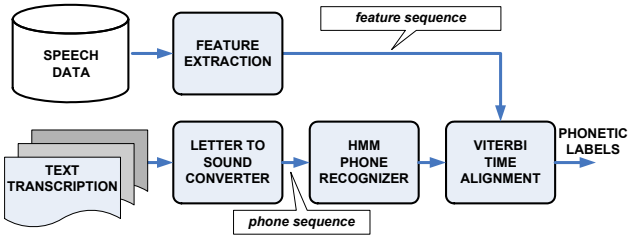

Fig. 1. Baseline HMM-based APS block diagram.

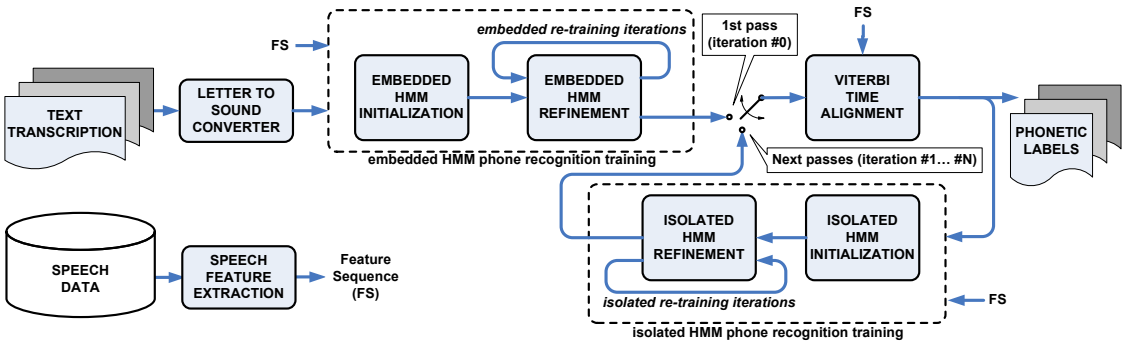

Fig. 2. Hybrid embedded-isolated phone alignment.

aligned against the corresponding phonetic transcriptions using the Viterbi algorithm [15].

In the present work we propose a hybrid architecture for automatic phone alignment which does not require knowledge of phoneme boundary information. Our approach combines the cost efficiency of embedded training techniques and the high performance of isolated-unit training.

2. EXISTING METHODS FOR HMM-BASED PHONE MODEL TRAINING

The training of each HMM-based phone model consists of two basic steps, initialization and refinement. Initially, prototype HMMs, corresponding to each phone, are defined and their parameters are initialized. Next, the parameters of each HMM are iteratively re-estimated in order to capture the corresponding phone's statistical characteristics. The Viterbi and Baum-Welch algorithms have been successfully used for that purpose. Typically, phone models are built from bootstrap speech data by employing either isolated-unit or embedded-unit training [16]. In isolated-unit training, hand-labeled bootstrap data are needed. Each HMM model is initialized and iteratively re-trained exclusively from the speech segments of the corresponding phone and the HMMs are trained independently from each other.

When labeled data are not available embedded-unit training is applied. Embedded training does not require any prior knowledge of the phone boundaries for the bootstrap data set. During initialization the training data are uniformly segmented and the parameters of every model are set equal to global values (flat initialization). Next, all HMMs are simultaneously re-estimated through the Baum-Welch algorithm, to update the parameters of the models.

It is known from the literature that isolated models achieve higher phoneme segmentation scores than embedded training [13]. However, when one deals with speech databases collected for the needs of specific application, often bootstrap data with hand-labeled phonetic transcriptions and time-marks of their boundaries are not available. In such cases, the dilemma is either to manually annotate part of the database (in order to prepare the bootstrap data) or to rely on embedded techniques to train the phone models.

3. THE HYBRID ARCHITECTURE

One problem that arises when embedded training is employed is the appearance of convergence problems for long sentences [10]. These problems are associated with the use of flat initialization. To avoid such inconvenience, we developed the hybrid architecture illustrated in Fig. 2. It takes advantage of the capability of embedded techniques to train HMMs without requiring information about the boundaries between the distinct units (here phones). On the other hand, it exploits the capability of isolated training to model more accurately (in the terms of maximum likelihood) the statistical characteristics of the target unit (phone).

As Fig. 2 illustrates, initially the word transcriptions of the bootstrap speech data are converted to the corresponding phone sequence. Since phone boundary information for the training data is not available, the embedded technique is exploited. One HMM is constructed for each phone and flat initialization is applied to every model. Subsequently, embedded re-estimation of the initialized models' parameters is performed. The training is terminated when the log-likelihood ratio between two successive iterations reaches a pre-defined threshold. After the refinement of the HMM parameters is completed, the speech data are force-aligned against the corresponding phone sequence. The outcome of the embedded HMM training procedure is an initial set of automatically estimated phone-labels. These phone-labels are fed as input to isolated-phone training. At first, a new set of models for the target phones is constructed. The new HMMs are initialized utilizing the automatic phone labels and further re-estimated. Similarly to the embedded training stage, a convergence criterion is introduced to control the number of re-estimation iterations. After refining the models' parameters is completed, the isolated-HMM phone models are utilized for re-alignment of the speech waveforms with the corresponding phone sequences. As a result, updated phone-labels are created. They are utilized as a feedback to construct new isolated HMM models, which subsequently will be time-aligned.

The process described so far, involving both embedded and isolated techniques, leads to automatically estimated phone labels, whose annotations are refined iteratively. The training process can be terminated when the overall bound-

ary shift between two successive iterations reach a predefined threshold. This hybrid architecture can be applied directly for training and time-alignment of speech data, or alternatively, for training HMM-based phone models from a bootstrap subset and then exploit them to segment speech data.

4. EXPERIMENTAL SETUP

To evaluate the proposed hybrid architecture we utilized the HTK toolkit [16] and the DARPA-TIMIT [17] database. TIMIT database has been established for measuring performance on the phone segmentation task [8, 11, 13].

Each utterance of TIMIT had been automatically segmented and manually checked. The labels correspond to the 61 phone set of American-English. Although in previous work [18] a subset of 48 phones has been proposed, here we use the full set in order to capture all phonetic phenomena that appear in speech.

For the purpose of comparison to previous related work [13], we followed the speech pre-processing and parameterization as described there. Specifically, speech waveforms were frame blocked every 5 milliseconds, utilizing a 20 milliseconds Hamming window. Employing a first-order FIR filter, pre-emphasis with factor equal to 0.97 was performed. For every speech frame, we computed the 12 first Mel frequency cepstral coefficients [16] and the 0-th cepstral coefficient, and their time derivatives. Thus, we consider a feature vector composed of 26 parameters.

In [13, 19], an HMM prototype with 6 states (the first and last state are non-emitting) and left-to-right transitions without skips was proposed. It was found to provide superior phone recognition rate. However, the full set of 61 phones contains also some shorter phones, whose duration is less than 20 milliseconds. In order to capture phones with duration close to 15 milliseconds we also consider a 5 state HMM, with two non-emitting states, left-to-right transitions without skips.

It has been shown in the literature that for the task of phone segmentation context-independent HMMs achieve higher segmentation accuracy [7, 20]. Therefore in the present evaluation, we consider context-independent HMMs.

5. EXPERIMENTS AND RESULTS

In all experiments, we followed the standard TIMIT division to training and test subsets. Thus, all models were created using the training subset as bootstrap data. However, these models were tested on both training and test subsets, in order to examine the performance on phone-aligning speech data as well as on building models for segmentation. Accuracy was evaluated in terms of percentage of boundaries predicted in location smaller than t milliseconds from the hand-labeled boundaries. The most commonly reported tolerance of 20 milliseconds was followed here [7].

As a first step, the phone segmentation accuracy for the case of embedded training (initial set of automatic phone-labels), as well as for the case of isolated training (using the hand-labeled transcriptions) were measured. The results obtained for the 5-state and 6-state HMMs are presented in Table 1. Here the **Train** and **Test** columns indicate the subset on which the performance was measured.

Table 1. Phone segmentation accuracy of embedded and isolated training using the hand-labeled phonetic transcriptions.

HMM	Training Method	Train	Test
5-states	Embedded	76.77%	76.31%
5-states	Isolated	87.47%	86.79%
6-states	Embedded	78.05%	77.47%
6-states	Isolated	88.67%	88.22%

The phone segmentation accuracy we obtained for the 6-state HMM model is in agreement with the performance reported in [13]. Furthermore, even on the complete set of 61 phones the 6-state HMM outperformed the 5-state one.

The accuracy of the automatic phone labels after each iteration, for the 6-state HMM, is illustrated in Fig. 3. As it can be seen in the figure, after a sufficient number of iterations, segmentation accuracy reaches 83.56% on the test subset. This corresponds to improvement of more than 6% in the accuracy of phone boundaries detection, when comparing to the baseline of 77.47% (for 0 isolated training passes). For the case of time-aligning of the training data, a segmentation accuracy of 84.03% was achieved.

Similar improvement of the phone segmentation rate was observed in the case of 5-state HMM. The accuracy of the automatic phone labels after each iteration, for the 5-state HMM, is illustrated in Fig. 4. As it can be seen in the figure, the accuracy increases from 76.31% (for 0 isolated training passes) to 82.61% after the first 20 iterations. When time-aligning of the training data was performed, the phone segmentation rate reached 83.12%, 6.35% higher than the baseline performance.

This significant improvement is owed to the iterative refinement of the phone models. Specifically, the gradual improvement of the detected boundaries leads to training of

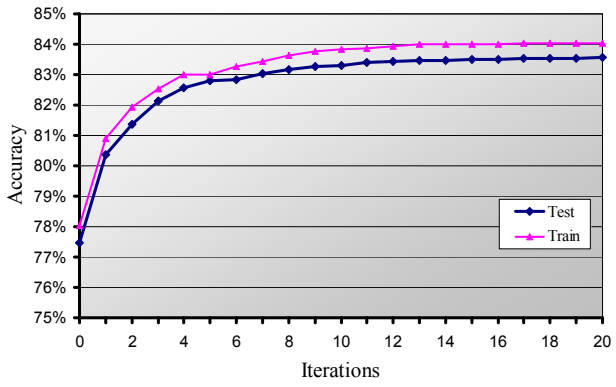

Fig. 3. Phone segmentation accuracy for the first 20 iterations and 6-state HMM phone models.

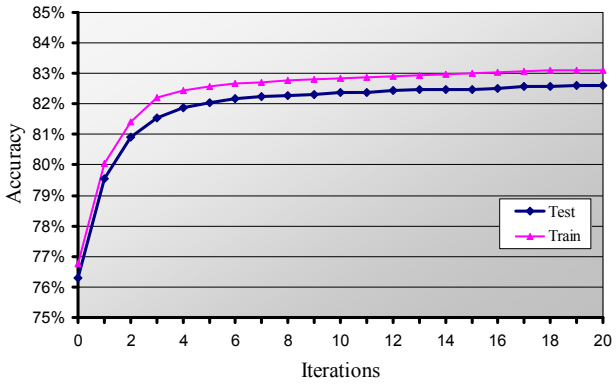

Fig. 4. Phone segmentation accuracy for the first 20 iterations and 5-state HMM phone models.

each phone model with more accurate annotations of the target speech segments (the bootstrap data that correspond to the each specific phone). In turn, more robust phone models lead to more accurate time-alignment of the speech waveforms and phonetic labels.

For the purpose of comparison with previous related work on TIMIT database, in Table 2 we present reported segmentation rates. The tabulated performances refer to the case of training without utilizing bootstrap data boundary information, where our method falls in.

Table 2. Reported phone segmentation accuracy on TIMIT.

Reported methods	Tolerance	Accuracy
Hybrid (present work)	<20 ms	83.60%
Brugnara et al. [13]	<20 ms	77.60%
Malfrere et al. [10]	<20 ms	80.21%
Ljolje et al. [21]	<17 ms	80.00%

6. CONCLUSION

We presented a hybrid embedded-isolated architecture for automatic time-alignment of speech waveforms and phone labels. Our approach does not require any phone boundary information, but offers advantage over the baseline embedded method. A 6% reduction of error from misaligned boundaries (for tolerance of 20 milliseconds), was achieved. We deem that the proposed method will benefit applications for which phonetic transcriptions are not available.

7. ACKNOWLEDGEMENT

This work was partially supported by the LOGOS project (ΕΗΓ-102), which is funded by the General Secretariat for Research and Technology of the Greek Ministry of Development.

8. REFERENCES

[1] J. Campbell, "Phoenetic, Idiolectal and Acoustical Speaker Recognition: Getting to Know you", CAIP Seminars, Nov. 5, 2003.

[2] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers, 1997.

[3] L.M. Arslan, "Speaker Transformation Algorithm Using Segmental Codebooks (STASC)", *Speech Communication*, vol.28, no.3, pp.211-226, 1999.

[4] T. Matsui, S. Furui, "Concatenated Phoneme Models for Text-Variable Speaker Recognition", *Proc of the ICASSP'93*, USA, vol.2, pp.391-394, 1993.

[5] B. Wildermoth, K.K. Paliwal, "Speaker Recognition Using Acoustically Derived Units", *Proc. Microelectronic Engineering Research Conference*, Brisbane, Australia, Nov. 2005.

[6] M.A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 1, pp.31-44, 1996.

[7] D.T. Toledano, L.A.H. Gomez, L.V. Grande, "Automatic Phonetic Segmentation", *IEEE Trans. on Speech and Audio Processing*, vol.11, no. 6, pp.617-625, 2003.

[8] B.L. Pellom and J.H. Hansen, "Automatic Segmentation of Speech Recorded in Unknown Noisy Channel Characteristics", *Speech Communication*, vol.25, pp.97-116, 1998.

[9] J.P. van Hemert, "Automatic Segmentation of Speech", *IEEE Trans. on Signal Processing*, vol. 39, no. 4, pp.1008-1012, 1991.

[10] F. Malfrere, O. Deroo, T. Dutoit and C. Ris, "Phonetic Alignment: Speech Synthesis-based vs. Viterbi-based", *Speech Communication*, vol.40, pp.503-515, 2003.

[11] J. Keshet, S.S. Shwartz, Y. Singer, D. Chazan, "Phoneme Alignment Based on Discriminative Learning", *Proc. of the Interspeech'05*, pp.2961-2964, 2005.

[12] K. Torkkola, "Automatic Alignment of Speech With Phonetic Transcription in Real Time", *Proc. of the ICASSP'98*, pp.611-614, 1998.

[13] F. Brugnara, D. Falavigna and M. Omologo, "Automatic Segmentation and Labeling of Speech Based on Hidden Markov Models", *Speech Communication*, vol.12, pp.357-370, 1993.

[14] J. Adell, A. Bonafonte, J.A. Gomez, M.J. Castro, "Comparative Study of Automatic Phone Segmentation Methods for TTS", *Proc. of ICASSP'05*, pp.309-312, 2005.

[15] J. D. Forney, "The Viterbi Algorithm", *Proc of the IEEE*, vol. 61, no. 3, pp. 268-278, 1978.

[16] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, "The HTK Book", (for HTK Version 3.3), Cambridge University, April 2005.

[17] J. Garofolo, "Getting Started with the DARPA-TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology (NIST), Gaithersburgh, MD, USA,1988.

[18] K.F. Lee, H.W. Hon, "Speaker-Independent Phone Recognition Using Hidden Markov Models", *IEEE Trans. on Acoustics Speech and Signal Processing*, vol.37, no.11, pp.1641-1648 1989.

[19] B.L. Pellom, J.H.L. Hansen, "Automatic Segmentation and Labeling of Speech Recorded in Unknown Noisy Channel Environments", *Proc. of the ESCA-NATO Workshop on Robust Speech Recognition for Unknown Communication Channels*, pp.167-170, 1997.

[20] A. Ljolje, M.D. Riley, *Automatic Speech Segmentation for Concatenative Inventory Selection*, Progress in Speech Synthesis, Springer, pp.305-311, 1997.

[21] A. Ljolje, M.D. Riley, "Aytomatic Segmentation and Labeling of Speech",*Proc. of the ICASSP'91*, pp.473-476, 1991.