Phonetic Segmentation of Emotional Speech with HMM-based Methods

Iosif Mporas, Todor Ganchev, Nikos Fakotakis

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

Abstract

In the present work we address the problem of phonetic segmentation of emotional speech. Investigating various traditional and recent HMM-based methods for speech segmentation, which we elaborated for the specifics of emotional speech segmentation, we demonstrate that the HMM-based method with hybrid embedded-isolated training offers advantageous segmentation accuracy, when compared to other HMM-based models used so far. The increased precision of the segmentation is consequence of the iterative training process employed in the hybrid-training method, which refines the model parameters and the estimated phonetic boundaries taking advantage of the estimations made at previous iterations. Furthermore, we demonstrate the benefits of using purposely-built models for each target category of emotional speech, when compared to the case of one common model built solely from neutral speech. This advantage, in terms of segmentation accuracy, justifies the effort for creating and employing the purposely-built segmentation models per emotional category, since it significantly improves the overall segmentation accuracy.

Keywords: Phonetic segmentation, hidden Markov models, emotional speech.

1. Introduction

Over the last years, there is an extensive use of automated systems, supporting voice or multimodal human-machine interaction [1], such as voice portals, call centers, e-banking, info kiosks, web services and applications etc. Due to the wide-spread use of this technology and the demand for convenient and efficient interaction, the design and development of natural and user-friendly speech interfaces became of primary importance.

In general, humans feel more natural when communicating with other humans because of the extra information represented in their non-verbal expressions can be recognized, processed, and reflected [2]. During a human-to-human interaction, there are two channels transmitting in parallel, one conveying

the explicit information of the uttered message and one transmitting implicit information about the speakers themselves [3, 4]. Due to the deficiencies of the present-day human-machine interaction technology, and specifically, due to the lack of machine-based emotional intelligence [2], there is a gap between the information conveyed and the information perceived when humans interact with automatic systems. Thus, one particular prerequisite for achieving naturalness of the human-machine interaction is the processing and comprehension of the implicit emotional information on the machine side, together with the explicit message.

Speech front-ends, which support emotional expressivity, include emotion recognition (ER) [ER, ER2], automatic speech recognition (ASR) of emotional speech and emotional text-to-speech (TTS) conversion. In general, the ASR is designed with one acoustic model [5] or language model [6] for each emotion of interest, while the TTS consists of one voice for each emotion [7, 8] or modulation of one voice to the specifics of each emotion [9, 10]. The performance of the speech front-end directly affects the naturalness of the human-machine interaction, i.e. the ER and ASR are responsible for the recognition of the incoming implicit and explicit information and the TTS is responsible for the feedback to the user and the expression of emotions.

Generally, the training of ER and ASR modules does not require phonetic time-alignment of the training speech recordings, since embedded algorithms [11] are used. Still, there are some approaches that exploit such information [12]. However, for the creation of a TTS voice, where the unit-selection approach has dominated the speech synthesis area, the availability of phonetic transition positions is a prerequisite. Moreover, since in the unit-selection approach the synthetic speech signal results from the concatenation of the speech units which are extracted from a training corpus, the precision of the time-alignment of the phonetic boundaries is essential for the quality of the synthetic speech signal.

The construction of an emotional TTS requires the availability of recordings of a speaker, together with the corresponding phonetic time-alignment, for each emotional speaking style of interest [7, 8]. Presently, the most precise way to annotate the positions of the phonetic transitions is manually by expert phoneticians. However, since manual phonetic time-alignment is a tedious, time-consuming and expensive task, automatic segmentation methods are usually employed. Presently, automatic segmentation methods achieve lower performance, comparing to the manual ones, thus corrections over the automatically extracted boundaries are usually performed by human annotators. In this connection, it is straightforward that the more accurate the automatic segmentation method is the less

time will be spent by the human annotator to correct the estimated phonetic boundaries. Particularly, in the case of direct use of the automatically estimated boundaries, without subsequent manual corrections, the misalignment of the phonetic boundaries has a significant effect on the quality of synthetic speech. Hence, the accuracy of the automatic segmentation is crucial in terms of time and cost demands as well as for the quality of the synthetic speech.

Several approaches for time-aligning speech waveforms with their corresponding phonetic sequences have been proposed, with most popular among them the dynamic time warping (DTW) [13] and the HMM-based [14-16] methods. In DTW, the original speech waveform is aligned against a synthetic speech signal, with known phone boundary positions, produced by an existing TTS. In the HMM-based approach the speech waveform is time-aligned against a phone label sequence utilizing a set of HMM phone models, i.e. a phone recognizer. The phone recognizer can be trained directly from the speech signals that are needed to be time-aligned, through the Baum-Welch algorithm [11].

The phonetic segmentation of an emotional speech database with the DTW approach would require the existence of one synthetic voice per emotion, which usually is not the case. Alternatively, the use of a non-emotional TTS would offer precise alignment only for emotions, whose speech characteristics are close to the ones of the neutral speaking style. On the other hand, the HMM-based approach offers the possibility to separately train HMMs for each emotion, and thus construct one phone recognizer dedicated to each category of emotional speech. Consequently, in the present work, due to the absence of an emotional TTS, we rely on the HMM-based approach, which we find more attractive for the task of phonetic segmentation of emotional speech.

In the task of emotional speech segmentation, a typical drawback for the HMM approach is the limited size of the available speech database [17]. Since there are some practical and ethical issues that restrict the possibilities for collecting real-life corpora of emotional speech, most of the existing emotional speech databases consist of limited-size acted speech, spoken from professional actors [18-21]. The usually restricted amount of data does not allow the training of robust models, which results to suboptimal phonetic segmentation accuracy.

In the past, few efforts have been spent on the task of emotional speech segmentation [17, 22]. These efforts have mainly focused on the utilization of the standard HMM-based approach, which is typically used for the segmentation of non-emotional speech, as well as on the possibility of merging training data from different emotion categories, in order to improve the robustness of the HMM

models, and therefore the precision of the boundary estimations.

In the present work, we count on a recently proposed HMM-based method for phonetic segmentation of speech, which employs hybrid embedded-isolated training [23], and we adapt it on the task of phonetic segmentation of emotional speech. This method, which in the following is referred to as *HYBRID* training method, is evaluated and its performance is contrasted against other HMM-based methods studied in the literature on the same task. In detail, we investigate the performance of the proposed method and compare it to a baseline HMM-based segmentation method, utilizing emotion-independent and emotion-dependent phone models. Additionally, in order to compensate for the limited amount of training data, we experiment with two model adaptation techniques for HMMs, which allow the model parameters of a general model to be adapted to the characteristics of the emotional speech for each category. Finally, we investigate the performance of the *HYBRID* training method for different types of training data and for the cross-category case of emotional speech segmentation, i.e. when the processed speech data do not match the category of emotional speech, for which the HMM model was built. None of the methods considered here makes use of bootstrap speech data with marked phonetic boundaries.

The remainder of this paper is organized as follows. In Section 2, we present the standard HMM-based method, which we consider here as the baseline as well as the *HYBRID* training method adapted for the task of emotional speech segmentation. In Section 3 we provide a description of the experimental setup, including the description of the emotional database and the settings of the HMMs. In Section 4, we present the experimental results and discuss various aspects of the experimental evaluation. Finally, Section 5 concludes this study.

2. HMM-based Segmentation of Emotional Speech

In contrast to non-emotional speech, where the spectral and prosodic (e.g. phone duration) characteristics of each phone exclusively depend on its context in a uniform manner, in emotional speech the phonetic characteristics present dissimilarities among the different emotions even for the same context. In phonetic terms, the variations in the phonetic characteristics can be attributed to differences in the voice quality, precision of articulation and deviations from the canonical segmental content [22]. Thus, the already tedious task of automatic phonetic segmentation becomes even more challenging in the case of emotional speech, since the spectral characteristics of each phone depend not

only on its phonetic context but also on the emotional state of the speaker. In the following subsections, we describe four HMM-based methods for automatic phonetic segmentation of emotional speech.

2.1. Baseline HMM-based Segmentation Method

The HMM-based segmentation method, which we consider here as the baseline, is inspired from the speech and phone recognition tasks and became popular because of its well known structure [ASR]. Specifically, in this method each emotional speech waveform is initially decomposed to a sequence of feature vectors, using a speech parameterization technique. Afterwards, a set of HMM phone models (phone recognizer) is utilized to extract the corresponding phonetic sequence as well as the positions of the phonetic boundaries. When segmenting speech corpora the word transcription of each speech waveform is usually known and can be converted to a phonetic label sequence through a letter-to-sound converter. In this linguistically constraint case, where the present work falls in, the phone recognizer is confined at the detection of the phonetic transition positions. Specifically, each phone label sequence is force-aligned against the corresponding feature vector sequence and the phone model set, through the Viterbi algorithm [24].

When segmenting emotional speech recordings, the utilized phone recognizer can be trained either from a merged dataset, which includes more than one categories of emotional speech – leading to emotion-independent (EI) phone models, or separately from speech data of the different categories of emotional speech – leading to emotion-dependent (ED) phone models. The block diagram of the HMM-based phonetic segmentation for emotional speech, in the linguistically constrained case is shown in Fig. 1, where the case of EI phone models is shown in Fig. 1(a), and the case of ED phone models in Fig. 1(b).

FIGURE 1

If preexisting phone recognizers are not available, phone models can be trained directly from the emotional speech data, which are going to be segmented (target data), with flat-initialization and parameter refinement via the Baum-Welch algorithm [11]. Alternatively, if manually segmented bootstrap speech data are available (which is not the case in the present study), they can be used for phone model training via the Viterbi algorithm [24].

2.2. Adapted HMM-based Segmentation Method

In the baseline HMM-based segmentation, discussed in Section 2.1, the training data of each category of emotional speech are utilized independently from the other categories, and are processed separately to train emotion-specific ED phone models. This splitting of the training data, in combination with the restricted size of the emotional speech database [18-21], does not allow the training of robust HMM models for each phone. One way to avoid this drawback is to design a two-step training process where at the first step we train EI phone models, and subsequently at the second step adapt them on the training dataset of each emotion. This method is described as HMM-based segmentation with ED model adaptation from a common EI model and in the present simply referred to as adapted HMM-based segmentation method.

In detail, initially Baum-Welch training is utilized to train HMM phone models over the entire speech data. The resulting emotion-independent HMMs are utilized as initial models from an adaptation algorithm, which adjusts the parameters of the HMMs on the speech dataset of each emotion. Several techniques for the adaptation of HMMs have been proposed, among which are the maximum likelihood linear regression (MLLR) [25] and the maximum a-posteriori (MAP) adaptation [26]. In the present work, we consider the MAP adaptation of the HMM phone models.

2.3. Hybrid HMM-based Segmentation Method

In [23], an HMM-based phonetic segmentation method that employs hybrid embedded-isolated training was introduced and successfully applied for phonetic segmentation of non-emotional speech. This method, which utilizes both Baum-Welch and Viterbi training to refine the HMMs parameters, was reported to offer superior segmentation performance, when compared to the baseline HMM-based method with embedded trained phone models.

The hybrid HMM-based segmentation method consists of two successive steps, the embedded initial training of the phone models and the isolated-unit iterative training. In detail, one HMM model for each phone is trained with flat-initialization of all the HMM parameters over a training speech data set and re-estimation of them with the Baum-Welch algorithm. The resultant HMM phone models are used to time-align the speech waveforms against their corresponding phonetic sequences and compute a first estimation of the positions of the phonetic boundaries. These boundaries are used for isolated-

unit (Viterbi) training of HMM phone models and subsequently for phonetic time-alignment. As a result, updated phone-boundaries are created, which are utilized as a feedback to construct new isolated HMM models, which subsequently will be time-aligned. After each iteration refined phone boundary positions are estimated and the iterative process terminates when the overall boundary shift between two successive iterations reach a predefined threshold. The advantage of the hybrid method comparing to the typical HMM segmentation is that enforces the training of the phone model parameters from speech samples of the specific phone and thus the Viterbi algorithm trains HMMs with sharper distributions, which leads to more accurate estimation of the phonetic transition positions.

In the present work, we study a modification of this method that situates it in compliance with the task of phonetic segmentation of emotional speech. In brief, the modifications, with respect to the scheme presented in [23] consist of the utilization of multiple phone recognizers, one for each emotional category, for the estimation of the initial phonetic boundaries of the emotional speech waveforms. Furthermore, after each isolated-unit training iteration one phone recognizer for each emotional category is constructed and separately used during Viterbi time-alignment. Thus, in contrast to the hybrid training architecture presented in [23], where in every step all the data were processed together and one refined phone recognizer was constructed, here each emotional training dataset is processed separately from the other datasets, and for each emotional category different phone recognizers are refined. The block diagram of the hybrid training method for phonetic segmentation of emotional speech is shown in Fig. 2.

FIGURE 2

In detail, the emotional speech recordings of interest are initially processed by an ED phone recognizer to time-align each speech waveform with the corresponding phonetic label sequence. After this initial phonetic boundary estimation, Viterbi training on the detected boundaries, followed by time-alignment using the new phone models is iteratively applied on the emotional speech recordings. After each iteration, more precise phone boundary positions are detected. After a sufficient number of iterations, the estimated phonetic transition positions do not change noticeably, and thus no further improvement in the segmentation accuracy is achieved. The training process is terminated when the overall boundary shift between two successive iterations reaches a predefined threshold. Thus, as

shown in Fig. 2, the final boundary estimations are the outcome of that Viterbi time-alignment, for which the termination criterion was met.

3. Experimental Setup

The performance of the HMM-based methods described so far was tested in the task of phonetic segmentation of emotional speech. Different HMM setups for the segmentation methods described in Section 2 were examined in a common experimental protocol.

3.1. Speech Database

In the present evaluation, we employed the Greek Emotional Speech (GrES) database [18]. The GrES database consists of recordings of a thirty-years-old Greek professional actress. It has been recorded in studio environment using a high-quality microphone, sampling frequency 44.1 kHz and resolution 16 bits per speech sample.

The recordings include acted speech of the five major emotion categories, i.e. *Anger*, *Fear*, *Joy*, *Neutral* and *Sad*. The same linguistically and prosodically rich text was uttered across all emotional speaking styles. The prompts, the meaning of which does not underlie any emotional concept, were extracted from the literature, newspapers and/or were set up by a professional linguist. All speech recordings have been annotated manually to their corresponding phonetic units by an expert phonetician. The data are distinguished to 25 short sentences, 21 long sentences and 16 paragraphs per emotion type, resulting in total to 310 speech files. Further information concerning the GrES database is available in [18].

3.2. Phone Recognizer

For the construction of the phone models, we utilized the HTK toolkit [27]. All phonetic models were trained from the corresponding speech data to be segmented, i.e. the training and test data fully overlapped.

For the needs of our evaluation all speech recordings were down-sampled to 16 kHz. The speech waveforms were frame blocked with a shifting Hamming window of 20 milliseconds, moving with step 5 milliseconds. Pre-emphasis with factor equal to 0.97 was performed, employing a first-order FIR filter. For every speech frame, we computed the 12 first Mel frequency cepstral coefficients [27] and

the 0-th cepstral coefficient. The delta and double-delta coefficients of the 13 static MFCC parameters referred above where appended to the initial feature vector constructing the final feature vector of length 39.

All words in the annotations of the GrES database were converted to their corresponding phone sequences, utilizing a set of 35 phones. This phone set is a modification of the SAMPA [28] alphabet for the Greek language. For each of the 35 phones and for each of the HMM methods outlined in Section 2, a left-to-right HMM, with three-states and no skipping transitions, was utilized. The states of the HMMs were modeled by one up to eight continuous Gaussian distributions. All phone models were context-independent, since they were observed to achieve higher accuracy in the task of phonetic segmentation [29, 30].

3.3. Experimental Protocol

For each of the HMM-based methods described in Section 2 a common experimental protocol was followed. In detail, we used the whole database to train EI phone models. Apart from these models, we also trained ED phone models, utilizing the corresponding data of each emotion, with Baum-Welch training (maximum likelihood criterion – ML). As an alternative to the ML training, the EI phone models were adapted with the speech recordings of each emotion using the MAP criterion. Finally, separately for each subset of emotional speech data, the ED phone models were used to obtain initial estimates of the phonetic transition positions and afterwards HMM phone models are trained through the iterative hybrid training method.

For all segmentation methods evaluated here, the data subsets used to train each phone recognizer were afterwards used as test data to measure the segmentation accuracy for each category of emotional speech. The only exception is the case of EI phone models where the training set consisted of recordings from all categories of emotional speech, and where the segmentation accuracy was estimated separately for each category as before.

The segmentation accuracy was measured in terms of the percentage of the estimated boundaries, whose misalignment is within the admissible tolerance of 20 milliseconds from the manually annotated boundary labels, which is the most commonly used figure of merit [14, 15, 23, 31]. This tolerance is considered as an acceptable limit for producing good quality synthetic speech [32, 33]. In addition, we report the segmentation accuracy in terms of mean absolute error (MAE) in milliseconds.

4. Experimental Results

In the experimental evaluation, we firstly investigate the segmentation accuracy for the phonetic segmentation methods outlined in Section 2. Afterwards using the best-performing method, we study various aspects of phone model creation and utilization.

4.1 Phonetic Segmentation Accuracy for the Different HMM-based Methods

The experimental results for the phonetic segmentation methods outlined in Section 2 are shown in Tables 1 and 2, in terms of MAE and in terms of segmentation accuracy for boundary misalignment within the tolerance of 20 milliseconds, respectively. The first column in the table indicates the number of mixture components *m* in each HMM state. In the tables, *EI* stands for the emotion-independent models, *ED* for the emotion-dependent models, *MAP* for the adapted models, and finally *HYBRID* for the phone models trained with the hybrid method. The best HMM setup, i.e. the most favorable number of mixtures per state, for each method and for each emotion is indicated in bold.

Table 1

Table 2

4.1.1 Influence of the method

As can be seen in Table 2, the *HYBRID* method achieves the highest segmentation accuracy, both in terms of MAE and in terms of misalignment within a tolerance of 20 milliseconds, for all categories of emotional speech. The experimental results, both in terms of MAE and misalignment within the tolerance of 20 milliseconds, indicate that the *MAP* and *ED* methods offer equivalent speech segmentation accuracy across the five evaluated emotional speaking styles. In addition, the *ED* phone models do not perform significantly better than the *EI* ones, which is in agreement with [22], were it was shown that the use of emotion-specific models do not significantly improve the overall performance and thus are not worth the effort. Exception is the *Neutral* case, where the use of ED models offers an improvement of the segmentation accuracy by more than 6%, in terms of misalignment within the admissible tolerance of 20 milliseconds, and by approximately two milliseconds reduction of the misalignment in terms of MAE.

As Table 1 presents, for the common EI acoustic model there are several HMM setups (e.g. 1, 3, 6 and 7 mixture components) for category *Joy*, where the MAE increases dramatically. Analysis of the errors in these specific setups found out that the significant increase of MAE, is due to the large number of gross-error misalignments, which are mainly observed at the beginning and at the end of words. The largest contributors for the high MAEs are the phonetic categories *stops, fricatives* and *vowels* at the beginnings and endings of words and to some extent *nasals* in the beginning of words. In general, the phone-to-phone transitions, where the phonetic categories *stops*, *fricatives* and *vowels* are involved, were found as the most affected by the acoustic mismatch between emotional speech from category *Joy* and the common EI acoustic model, which is general enough not to model precisely *Joy*.

4.1.2 Influence of the emotion type

In the following, we focus out attention to the best-performing segmentation method, *HYBRID*, and its segmentation accuracy for the different categories of emotional speech. As the experimental results presented in Tables 1 and 2 show, there are significant differences in the segmentation accuracy among the different categories of emotional speech. The highest phonetic segmentation accuracy, in terms of admissible misalignment within tolerance $t \leq 20$ms, was observed for the category *Joy*, while the lowest one for *Anger*. Specifically, in the case of *Joy*, which is a relatively modal voice, the *HYBRID* method achieved 84.2%. Following, relatively high segmentation accuracy was observed for the category *Neutral* speech, which does not present extreme speaking modes or other emotive bursts, and for the category *Fear*, which is soft speech of low intensity. However, the *HYBRID* method offered significantly lower phonetic segmentation accuracy – 75.8 % and 77.9 % for the categories *Anger* and *Sad*, respectively. The relatively lower accuracy for category *Anger* that is observed here is in agreement with [22], and is mainly owed to the flustered speech with weakly articulated structures, which often results to weak or missing occlusions in stop sounds or even the omission of whole segments. The difficulty in segmenting speech from the category *Sad* is owed to the presence of stuttering and quite breathy voice, which does not facilitate the segmentation process. Another drawback in the phonetic segmentation performance of *Sad* speech could be the presence of not-loud speech, which results to not clear articulation of the corresponding phones.

4.1.3 Influence of the number of mixtures in the HMM setup

Another interesting observation is that for any of the methods, there is no specific HMM setup (i.e. number of Gaussian mixtures per state) that presents superior performance across all emotional categories. This behavior is owed to the different characteristics of the emotional categories, with respect to the clarity of the produced spectrum, the insertion of laughter and breath, and the precision of the articulation. The experimental results presented in Table 1, in terms of MAE and for the best performing *HYBRID* method, indicate that for the categories *Fear*, *Joy* and *Sad*, the phonetic segmentation accuracy is higher for the HMMs with many Gaussian mixtures per state than for the ones with few. For the category *Anger*, the performance is better when fewer mixtures are used, while in the category *Neutral*, the effect of the number of mixtures per HMM state affects significantly less the segmentation accuracy.

In previous studies on phonetic segmentation of non-emotional (neutral) speech [31, 34] it was shown that HMMs with fewer Gaussian components achieve higher segmentation accuracy, due to the inherent variance of the spectrum in the vicinity of a phonetic transition, which makes a simpler model more adequate. However, for the *HYBRID* method this variation in the spectral characteristics and the manner of articulation characteristic for the emotional categories *Fear*, *Joy* and *Sad*, when modelled by several rather than a single Gaussian distribution contributes for improving the phonetic segmentation accuracy.

As Table 1 presents, the number of mixture components in the some setups were either too many, such as in *Joy* for ED and MAP models for eight mixtures, or too few, such as in *Neutral* and *Sad* for EI models and one mixture. These results indicate the simultaneous dependency of the segmentation accuracy on both the amount of available training data and the underlying distribution of the data for each emotion category. The last well explains the observation that different HMM setups were found out as optimal for segmentation of the different emotion categories, even for the same speech segmentation method.

4.1.4 MAE vs. admissible tolerance of misalignment

As seen in Tables 1 and 2 for the best performing *HYBRID* method, there is some discrepancy between the performance estimation in terms of MAE and in terms of segmentation accuracy for misalignment within the admissible tolerance $t \leq 20$ milliseconds. This disagreement between the two figures of merit

is owed to the fact that the two criteria take into account different aspects of the segmentation accuracy. Indeed, the MAE shows the overall misalignment error including the gross errors [35], while the estimated segmentation accuracy for misalignment within a given tolerance only counts in boundary estimations, for which the error does not exceed the admissible tolerance. Thus, gross errors do not significantly affect the segmentation performance, when it is presented in terms of segmentation accuracy within an admissible tolerance.

From a practical point of view, the difference between the two figures of merit used here is more clear in the case of EI models for *Joy* and 1, 3, 6 and 7 mixtures per state, for *Neutral* and *Sad* for 1 mixture, as well as for ED and MAP models in *Joy* with 8 mixtures. In these cases, as presented in Table 1, the MAE is notably high, while the segmentation accuracy for misalignment within the admissible tolerance $t \leq 20$ milliseconds (Table 2) does not drop significantly. This difference indicates that for the specified cases many of the misalignments outside the admissible tolerance were in fact gross errors.

From the results discussed in this section, we can conclude that the hybrid training for HMM-based phonetic segmentation of emotional speech is the most appropriate method when compared to other HMM-based methods, such as HMMs with emotion-independent, emotion-dependent or emotion-adapted phone models.

4.2 Phonetic Segmentation Accuracy for Different Training Data Types

As a next step, we examined the effect of the utterance length in the training data (i.e. short sentences, long sentences and paragraphs) on the phonetic segmentation accuracy and the cross-emotional phonetic segmentation accuracy of models trained on data from different emotional category than the test recordings. The effect of these two factors was also examined in [17], where they were found to affect the accuracy of the segmentation of emotional speech. The phonetic segmentation performance was examined utilizing the best-performing method, i.e. the *HYBRID* method. For each category of emotional speech, the most successful HMM setup in terms of segmentation performance (in bold in Tables 1 and 2) was selected. According to the results shown in these tables, for the segmentation accuracy presented in terms of *MAE*, we selected the HMM models with number of mixtures per state m=2, 6, 5, 6, 8 for the emotions *Anger*, *Fear*, *Joy*, *Neutral* and *Sad*, respectively. Likewise, in terms of

segmentation accuracy for misalignment within the admissible tolerance $t \leq 20$ms, we selected these with m=4, 6, 2, 1, 3. In the following we discuss only the selected results.

In Table 3 we present the phonetic segmentation performance obtained when training the HMM models with training data of different sentence lengths, i.e. short sentences, long sentences and paragraphs, tested on the corresponding data types. The reported segmentation accuracies, in terms of MAE, correspond to measurements of the performance only on the corresponding types of training data. The colored cells correspond to the training set types that achieved higher segmentation accuracy than the merged training dataset, denoted as 'All data'. The best segmentation accuracy for each emotion is indicated in bold. Similarly, Table 4 presents the phonetic segmentation accuracy in terms of misalignment within the admissible tolerance $t \leq 20$ milliseconds for different types of training data, with respect to the sentence length.

Table 3

Table 4

As can be seen in Tables 3 and 4, the use of the entire training set, 'All data', for the training of the phone models does not always improve the overall phonetic segmentation performance. In detail, in terms of MAE, merging all speech data to a common training set decreases the performance for all the evaluated emotion categories and for most of the training set types. As presented in Table 4, for most training sets, the segmentation accuracy in terms of admissible misalignment tolerance $t \leq 20$ milliseconds is higher for set-dependent training. Exception is the category *Anger*, where none of the set-specific training sets outperforms the merged training set. This observation is owed to the different way that a given speaker is expressing the same emotion in paragraphs and short sentences [17, 36], which results to different articulation of same phones, even in the same context. Thus, using data of different speaking styles does not help for the improvement of the phonetic segmentation accuracy. This is in agreement with [17], where similar results were found when using sentences and paragraphs.

In Tables 5 and 6, we present phonetic segmentation accuracies for the cross-emotional evaluation, in terms of MAE and misalignment within the admissible tolerance of $t \leq 20$ milliseconds, respectively. In both tables, the rows correspond to the set of speech data, which was used to train the HMM phone models, while the columns correspond to the speech data on which each phone model was evaluated.

Table 5

Table 6

The results (i.e. the segmentation accuracy both in terms of MAE and in terms of misalignment within the admissible tolerance $t \leq 20$), show that not always the phone models trained on data of a specific emotion category can best segment emotional speech of that category. For example, it was observed that emotional speech from the categories *Anger* and *Sad* is segmented with better accuracy by the phone models trained for *Joy*, than by the ones trained with speech data from their own category. This is in agreement with [17] where 'sadness' was found to be segmented best from phone models trained with 'happiness' recordings. Furthermore, the conclusion of Gallardo-Antolin et al. [17] that fast-speech models can better segment slow-speech than the other way around is partially confirmed here. For example, *Sad* is better segmented by *Fear* and *Joy* models, when compared to HMMs trained from the same emotional type. Oppositely, in the case of *Neutral* speech, the 'faster' *Fear* and *Joy* phone models do not improve the precision of the phonetic time-alignment.

Finally, we investigated the accuracy for cross-emotion segmentation when using a common number of mixtures per HMM state for all emotion categories, instead of the best performing model for each category as it was above. For that purpose, we selected the *HYBRID* method with six Gaussian components, which corresponds to the best setup in terms of MAE, and three mixtures for the case when segmentation accuracy is measured in terms of misalignment within the admissible tolerance $t \leq 20$ milliseconds. The corresponding results are shown in Tables 7 and 8, respectively.

Table 7

Table 8

As can be seen in Tables 7 and 8, even for a common number of mixtures per state, specific categories of emotional speech are not segmented best by the models trained with emotional speech of the same category. Similarly to the results in Tables 5 and 6, here we observe that emotional speech from the categories *Anger* and *Sad* is segmented with better accuracy (in terms of MAE) by the phone models trained for *Joy*, than by the ones trained with speech data from their own category. However, in

terms of misalignment within the admissible tolerance $t \leq 20$ milliseconds, only emotional speech from category *Anger* is not segmented best by its own model, but by the model of *Joy*.

In summary, the experimental results presented in this section show two different aspects that concern the automatic segmentation of emotional speech, namely the length of the spoken utterance and the appropriateness of phone models from specific emotion category to segment emotional speech from other categories. Particularly, our observations led us to the conclusion that data from specific categories of emotional speech could be merged together and utilized for training common HMM models. Such merging might facilitate the partial alleviation of the problem with the shortage of training data.

5. Conclusion

In the present work, we examined the performance of four HMM-based methods on the task of phonetic segmentation of emotional speech. In detail, the baseline HMM-based speech segmentation approach was evaluated using emotion-independent, emotion-dependent and emotion-adapted phone models, and compared against a recently proposed HMM method with hybrid embedded-isolated training, which was modified here for the needs of emotional speech segmentation. The experimental results demonstrated that the hybrid training method detects the phonetic transition positions of emotional speech significantly better than the other methods, for all categories of emotional speech considered here. Specifically, the hybrid training method improved the segmentation accuracy in terms of misalignment within the admissible tolerance of 20 milliseconds by approximately 7.6% for *Anger*, 19.1% for *Fear*, 18.5% for *Joy*, 8.1% for *Neutral* and 14.2% for *Sad*, when compared to the second best performing in each case EI/ED/MAP HMM-based method. These results confirm the advantage of the hybrid training method on the task of phonetic segmentation of emotional speech.

Furthermore, as reported in Section 4, the phone segmentation accuracy for emotional speech improves for matching train and test conditions, such as same sentence length, rhythm and speaking style. Thus, these factors should be taken into consideration, for optimizing the phonetic segmentation accuracy. Finally, in the cross-emotional experiments it was observed that phone models trained on speech data of specific emotional categories can offer a more accurate phonetic time-alignment, than others.

The task of emotional speech segmentation was observed as more tedious than the segmentation of neutral speech. This is owed to the presence of implicit information in the speech signal, which indicates the emotional state of the speaker, and results in a greater variation of the spectral characteristics of phones, even when they reside in the same phonetic context. This variability makes the detection of the phonetic transition positions more difficult, when emotional speech is segmented.

Acknowledgement

This work was partially supported by the PlayMancer project (FP7 215839), which is co-funded funded by the FP7 of the European Commission.

References

[1] Z.H. Zeng, J.L. Tu, M. Liu, T.S. Huang, B. Pianfetti, D. Roth, and S. Levinson, "Audio-visual affect recognition", *IEEE Transactions on Multimedia*, 9(2) 424-428, 2007.

[2] D. Morrison, R. Wang, and L.C. De Silva, "Ensemble methods for spoken emotion recognition in call-centres", *Speech Communication*, 49 (2007) 98-112.

[3] R. Cowie, and E. Douglas-Cowie, "Speakers and hearers are people: Reflections on speech deterioration as a consequence of acquired deafness," in *Profound Deafness and Speech Communication*, K-E. Spens and G. Plant, Eds. London, UK: Whurr, 1995, pp. 510-527.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, 18(1) 32-80, 2001.

[5] T. Kostoulas, I. Mporas, T. Ganchev and N. Fakotakis, "The effect of emotional speech on a smart home application", in *Lecture Notes in Computer Science*, Springer Berlin/ Heidelberg, 5027/2008, pp. 305-310.

[6] T. Athanaselis, S. Bakamidis, I. Dologlou, R. Cowie, E. Douglas-Cowie, and C. Cox, "ASR for emotional speech: Clarifying the issues and enhancing performance", *Neural Networks*, 18(4), 437-444, May 2005.

[7] M. Schröder, "Emotional speech synthesis - A review", in *Proc. of Eurospeech 2001*, vol. 1, pp. 561-564, 2001.

[8] J.M. Montero, J.M. Gutierrez-Arriola, S. Palazuelos, E. Enriquez, S. Aguilera, and J.M. Pardo, "Emotional speech synthesis: from speech database to TTS", in *Proc. of ICSLP 1998*, p.1037, 1998.

[9] P.-Y. Oudeyer, "The production and recognition of emotions in speech: features and algorithms", *International Journal of Human-Computer Studies*, 62(3), pp. 423, 2005.

[10] G.O. Hofer, K. Richmond, and R.A.J. Clark, "Informed blending of databases for emotional speech synthesis", in *Proc. of Interspeech 2005*, pp. 501-504, 2005.

[11] L.E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". *Annals of Mathematical Statistics*, 41(1), 164-171, 1970.

[12] V. Sethu, E. Ambikairajah, and J. Epps, "Phonetic and speaker variations in automatic emotion classification", in *Proc. of Interspeech 2008*, pp. 617-620, 2008.

[13] F. Malfrere, O. Deroo, T. Dutoit, C. Ris, "Phonetic alignment: speech synthesis-based vs. Viterbi-based", *Speech Communication*, 40 (2003) 503–515.

[14] B.L. Pellom, and J.H.L. Hansen, "Automatic segmentation of speech recorded in unknown noisy channel characteristics", *Speech Communication*, 25 (1998), 97–116.

[15] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden Markov models", *Speech Communication*, 12 (1993) 357–370.

[16] A. Ljolje, J. Hirschberg, J.P.H. Van Santen, "Automatic speech segmentation for concatenative inventory selection", in J.P.H. Van Santen, R.W. Sproat, J.P. Olive and J. Hirschberg, Editors, Progress in Speech Synthesis, Springer (1997), pp. 304–311.

[17] A. Gallardo-Antolín, R. Barra, M. Schröder, S. Krstulovic, and J.M. Montero, "Automatic phonetic segmentation of Spanish emotional speech", in *Proc. of Interspeech 2007*, pp. 2905-2908, 2007.

[18] P. Zervas, I. Geourga, N. Fakotakis, and G. Kokkinakis, "Greek emotional database: construction and linguistic analysis", in *Proc. of 6th International Conference of Greek Linguistics*, 2003.

[19] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts", Linguistic Data Consortium, Philadelphia, 2002.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, and B. Weiss, "A database of German emotional speech", in *Proc. of Interspeech 2005*, pp. 1517-1520, 2005.

[21] I.S. Engberg, and A.V. Hansen, "Documentation of the Danish emotional speech database (DES)", AAU report, Person Kommunication Center, Denmark, 1996.

[22] H. Pirker, "Phonetic segmentation of the GEMEP-corpus: Applying forced alignment to emotional speech", Technical Report, Austrian Institute for Artificial Intelligence, Wien, TR-2007-11, 2007.

[23] I. Mporas, T. Ganchev, and N. Fakotakis, "A hybrid architecture for automatic segmentation of speech waveforms", in *Proc. of ICASSP 2008*, pp. 4457-4460, 2008.

[24] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", *IEEE Trans. Information Theory*, 13(2) 260-269, 1967.

[25] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", *Computer, Speech and Language*, 9(2) 171-185, 1995.

[26] J.L. Gauvain, and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Transactions on Speech and Audio Processing*, 2(2) 291-298, 1994.

[27] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Cambridge University Engineering Department, 2006.

[28] J.C. Wells, "SAMPA computer readable phonetic alphabet", in Gibbon, D., Moore, R. and Winski, R. (eds.), 1997. Handbook of Standards and Resources for Spoken Language Systems. Berlin and New York: Mouton de Gruyter. Part IV, section B.

[29] D.T. Toledano, L.A.H. Gomez, and L.V. Grande, "Automatic phonetic segmentation", *IEEE Trans. on Speech and Audio Processing*, 11(6) 617-625, 2003.

[30] A. Ljolje, and M.D. Riley, "Automatic speech segmentation for concatenative inventory selection," *Progress in Speech Synthesis*, Springer, pp. 305-311, 1997.

[31] I. Mporas, T. Ganchev, and N. Fakotakis, "Speech segmentation using regression fusion of boundary predictions", *Computer, Speech and Language*, 24(2) 273-288, 2010.

[32] J. Matousek, D. Tihelka, and J. Psutka, "Automatic segmentation for Czech concatenative speech synthesis using statistical approach with boundary-specific correction", in *Proc. of Interspeech 2003*, pp. 301–304, 2003.

[33] L. Wang, Y. Zhao, M. Chu, J. Zhou, Z. Cao, "Refining segmental boundaries for TTS database using fine contextual-dependent boundary models". in *Proc. of ICASSP 2004*, vol. 1, pp. 641-644, 2004.

[34] D.T. Toledano, L.A.H. Gomez, and L.V. Grande, "Automatic phonetic segmentation", *IEEE Transactions on Speech and Audio Processing*, 11(6) 617–625, 2003.

[35] I.H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, San Francisco, 2005.

[36] X. Huang, A. Acero, and H.W. Hon, *Spoken language processing: A guide to theory, algorithm and system development*, Prentice Hall, Upper Saddle River, NJ, USA, 2001.

[ER] M. You, G.-Z. Li, J.Y. Yang, M.Q. Yang, "An enhanced lipschitz embedding classifier for multi-emotion speech analysis", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 8, pp. 1685-1700, 2009.

[ER2] T. Sobol-Shikler, P. Robinson, "Classification of Complex Information: Inference of Co-Occurring Affective States from Their Expressions in Speech", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Accepted for future publication.

[ASR] Q.H. He, S. Kwong, K.F. Man and K.S. Tang, "An improved maximum model distance approach for HMM-based speech recognition systems", *Pattern Recognition*, vol. 33, no. 10, pp. 1749-1758, 2000.

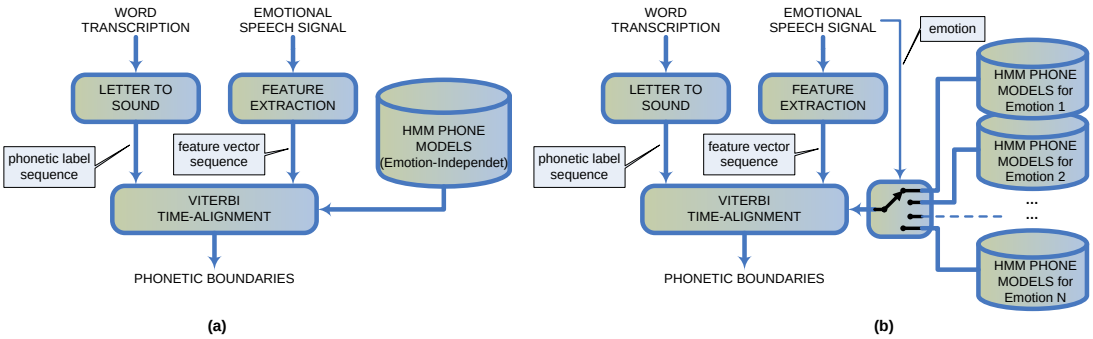

Fig. 1. Block diagram of the HMM-based phonetic time-alignment for emotional speech, utilizing (a) emotion-independent phone models and, (b) emotion-dependent phone models.

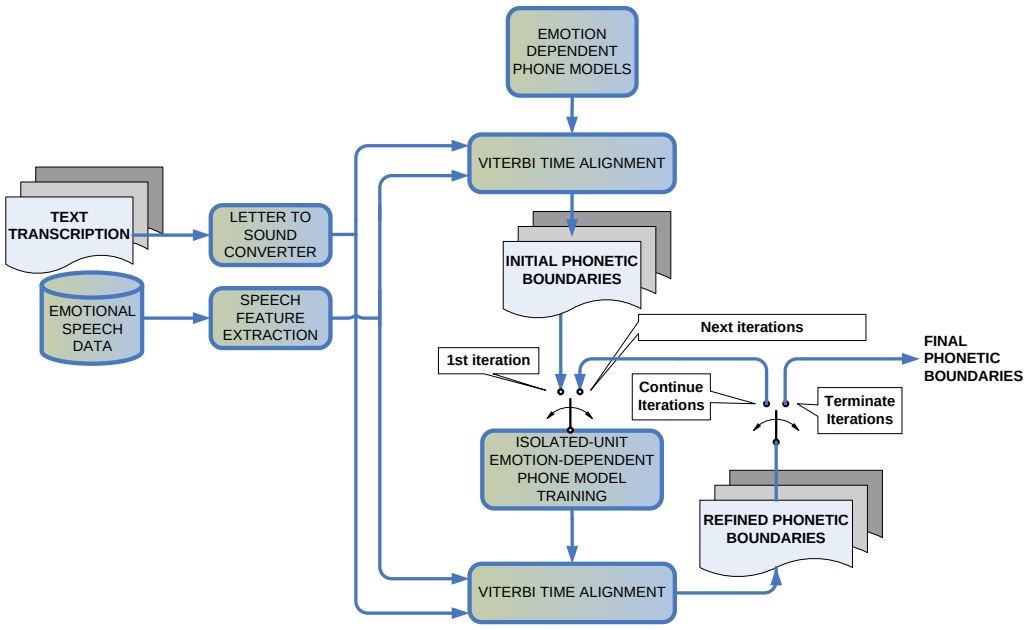

Fig. 2. Hybrid embedded-isolated phonetic alignment of emotional speech

Table 1. Segmentation performance in terms of mean absolute error (MAE). All numbers are the misalignment in milliseconds.

	Anger				Fear				Joy				Neutral				Sad			
m	**EI**	**ED**	**MAP**	**HYBRID**	**EI**	**ED**	**MAP**	**HYBRID**	**EI**	**ED**	**MAP**	**HYBRID**	**EI**	**ED**	**MAP**	**HYBRID**	**EI**	**ED**	**MAP**	**HYBRID**
1	50.3	23.0	23.1	17.8	53.1	25.1	24.8	19.3	149.1	28.5	28.5	18.0	95.6	19.8	19.6	14.1	134.7	41.2	39.8	28.8
2	34.8	17.8	17.8	**14.2**	39.7	19.7	19.8	16.0	24.5	21.4	21.4	15.7	34.4	17.8	**17.1**	14.5	21.4	23.5	23.5	17.0
3	37.6	17.4	18.0	15.3	44.7	20.4	20.1	13.2	108.8	23.5	22.8	16.2	32.7	18.6	18.3	**14.0**	49.8	21.7	21.6	15.8
4	37.3	16.6	16.6	14.3	45.0	20.2	20.2	13.7	32.2	20.8	20.7	16.1	**19.8**	24.1	18.0	14.1	23.7	20.6	20.5	14.4
5	16.2	19.3	17.9	16.9	19.5	23.1	20.7	12.3	**19.3**	**19.6**	**19.6**	**13.1**	20.0	18.4	18.8	19.9	**18.7**	19.0	18.9	43.9
6	**15.1**	**15.5**	**15.7**	16.9	**19.4**	**18.6**	**18.6**	**12.2**	126.9	21.4	21.3	13.5	20.1	**17.5**	17.4	**14.0**	**18.7**	**18.5**	**18.3**	14.7
7	15.5	16.4	16.0	16.7	19.5	19.3	18.9	12.4	116.5	20.9	20.8	15.8	23.2	17.7	17.6	15.2	18.8	18.8	18.9	42.9
8	15.8	16.9	16.9	17.2	19.7	19.2	19.2	13.0	19.6	110.9	114.8	14.4	23.1	18.3	24.3	14.9	19.2	19.1	19.2	**14.2**

Table 2. Segmentation accuracy for misalignment within the admissible tolerance of t≤20 milliseconds for the estimated boundaries. All numbers are in percentages.

	Anger				Fear				Joy				Neutral				Sad			
m	**EI**	**ED**	**MAP**	**HYBRID**	**EI**	**ED**	**MAP**	**HYBRID**	**EI**	**ED**	**MAP**	**HYBRID**	**EI**	**ED**	**MAP**	**HYBRID**	**EI**	**ED**	**MAP**	**HYBRID**
1	57.1	58.3	58.5	73.6	54.7	54.5	55.5	71.3	53.9	57.9	57.8	81.9	63.8	**74.0**	**74.6**	**82.7**	57.5	58.8	59.1	74.4
2	64.5	65.5	65.3	75.1	**59.5**	**61.9**	**61.3**	69.8	**64.9**	**65.7**	**65.8**	**84.2**	**67.4**	71.0	72.5	80.6	63.5	63.0	63.0	70.9
3	65.1	65.0	64.2	75.7	58.5	58.9	60.2	79.0	62.9	64.2	64.1	82.1	64.6	70.0	70.8	81.5	61.5	61.5	61.6	**77.9**
4	65.4	65.2	65.3	**75.8**	58.2	57.8	57.8	76.8	63.1	63.9	64.0	82.7	64.2	67.5	69.1	79.8	61.6	**63.3**	**63.2**	76.1
5	66.3	64.9	64.2	73.6	**59.5**	58.0	58.5	80.7	63.7	64.3	64.2	82.8	64.7	70.7	70.3	79.0	62.3	61.6	60.9	76.1
6	**68.2**	**66.5**	**66.3**	73.8	59.1	59.7	59.9	**81.0**	63.4	63.7	64.0	81.5	65.7	71.6	72.0	79.1	62.8	63.1	62.6	73.7
7	67.2	64.9	66.1	73.5	58.6	58.6	59.9	80.4	63.1	63.1	63.6	77.7	66.5	71.6	71.9	77.6	**63.7**	63.0	62.9	73.9
8	66.5	64.0	63.9	71.9	58.0	59.1	58.9	78.7	64.4	61.8	62.4	79.4	66.0	71.9	69.3	78.7	63.1	62.9	62.6	75.4

Table 3. Segmentation performance in terms of MAE (in milliseconds) for different training sets. Colored cells correspond to the training sets, which offered higher segmentation accuracy than the overall training set ('All data').

Training Set Type	Anger	Fear	Joy	Neutral	Sad
Short Sent.	14.6	13.4	21.3	13.5	15.9
Long Sent.	**13.9**	12.4	12.6	13.8	**13.7**
Paragraphs	16.3	**12.0**	**12.4**	**12.8**	14.1
All data	14.2	12.2	13.1	14.0	14.2

Table 4. Segmentation accuracy in percentages for misalignment within the admissible tolerance $t \leq 20$ milliseconds for different training sets. Colored cells correspond to the training sets, which offered higher segmentation accuracy than the overall training set ('All data').

Training Set Type	Anger	Fear	Joy	Neutral	Sad
Short Sent.	72.6	77.8	79.1	78.9	74.1
Long Sent.	73.7	80.5	84.0	84.4	**78.6**
Paragraphs	68.3	**81.6**	**84.3**	**85.2**	74.4
All data	**75.8**	81.0	84.2	82.7	77.9

Table 5. Cross-emotional segmentation accuracy in terms of MAE (in milliseconds) for the *HYBRID* method

	Test Set				
Training Set	**Anger**	**Fear**	**Joy**	**Neutral**	**Sad**
Anger	14.2	23.0	17.0	23.1	17.6
Fear	24.0	12.2	23.9	20.5	**14.2**
Joy	**11.0**	16.1	**13.1**	17.5	15.0
Neutral	24.3	**12.1**	58.0	**14.0**	16.8
Sad	16.1	13.0	19.3	24.7	**14.2**

Table 6. Cross-emotional segmentation accuracy in percentages for misalignment within the admissible tolerance t≤20 milliseconds for the *HYBRID* method

	Test Set				
Training Set	**Anger**	**Fear**	**Joy**	**Neutral**	**Sad**
Anger	75.8	66.0	76.8	68.2	71.2
Fear	59.5	**81.0**	71.2	72.7	78.1
Joy	**81.4**	74.8	**84.2**	74.0	**78.5**
Neutral	57.0	74.1	72.8	**82.7**	76.3
Sad	71.6	75.5	77.3	69.1	77.9

Table 7. Cross-emotional segmentation accuracy in terms of MAE (in milliseconds) for 6 mixtures per HMM state.

	Test Set				
Training Set	**Anger**	**Fear**	**Joy**	**Neutral**	**Sad**
Anger	16.9	24.6	16.3	23.9	17.6
Fear	24.0	12.2	23.9	20.5	**14.2**
Joy	**16.7**	16.3	**13.5**	17.6	**14.2**
Neutral	24.3	**12.1**	58.0	**14.0**	16.8
Sad	17.9	14.5	22.3	22.7	14.7

Table 8. Cross-emotional segmentation accuracy in percentages for misalignment within the admissible tolerance t≤20 milliseconds for 3 mixtures per HMM state

	Test Set				
Training Set	**Anger**	**Fear**	**Joy**	**Neutral**	**Sad**
Anger	75.8	63.3	76.1	64.3	69.4
Fear	61.5	**79.0**	73.3	71.9	77.3
Joy	**78.5**	73.1	**82.1**	72.8	76.4
Neutral	52.3	75.8	70.3	**81.5**	77.4
Sad	71.6	75.5	77.3	69.1	**77.9**