Unsupervised Speaker and Expression Factorization for Multi-Speaker Expressive Synthesis of Ebooks

Langzhou Chen, Norbert Braunschweiler

Toshiba Research Europe Ltd., Cambridge, UK
langzhou.chen,norbert.braunschweiler@crl.toshiba.co.uk

Abstract

This work aims to improve expressive speech synthesis of ebooks for multiple speakers by using training data from many audiobooks. Audiobooks contain a wide variety of expressive speaking styles which are often impractical to annotate. However, the speaker-expression factorization (SEF) framework, which has been proven to be a powerful tool in speaker and expression modelling usually requires the (supervised) information about expressions in the training data. This work presents an unsupervised SEF method which implements the SEF on unlabelled training data in the framework of cluster adaptive training (CAT). The proposed method integrates the expression clustering and parameter estimation in a single process to maximize the likelihood of the training data. Experimental results indicate that it outperforms the cascade system of expression clustering and supervised SEF, and significantly improves the expressiveness of the synthetic speech of different speakers.

Index Terms: expressive speech synthesis, hidden Markov model, cluster adaptive training, factorization, audiobook

1. Introduction

The modelling of speakers and expressions are important issues in TTS research. Given that the training data has discrete speaker and expression labels, various methods can be shared by these two research aspects, e.g., model interpolation [2, 3], transform based method [4, 5], CAT [6, 7]. Some of the methods which were proposed for speaker modelling, can be used for expression modelling as well, e.g., the eigenvoice method [8] and factor analyzed voice models [9]. The only difference is the definition of training data, i.e., speaker dependent data or expression dependent data.

All the methods mentioned above only deal with either speaker modelling or expression modelling. However, in some applications, the speaker and the expression need to be modelled simultaneously. For example, synthesizing an ebook may require the TTS system to read the stories expressively with the voices of different speakers. Directly modelling every combination of speaker and expression is often impractical since the expressive training data is not always available for every speaker. [10] presented a method to interpolate the neutral model of a target speaker with the emotional model selected from a candidate pool. However, interpolating the models from two different speakers will influence the similarity of the synthetic speech to the target speaker, especially when the coverage of the candidate pool is not big enough.

A better solution for this problem is achieved by factorization techniques which model speaker and expression independently when using training data with multiple expressions and speakers. This way, different speakers voices can share the speaker independent expressions and produce expressive synthetic speech. In [11], the CAT based speaker and expression factorization was presented. In [12], a speaker and language factorization (SLF) method was proposed which used CMLLR transforms to represent the speakers and the CAT weight vectors to represent languages. This can be extended to SEF by using CAT weight vectors to represent expressions rather than languages. The factorization methods mentioned above are based on labelled data, i.e. speaker and expression information in the training data is known.

Nowadays, huge amounts of audiobook data is available and has been used for TTS system training [13]. This data source contains highly diverse speech which covers a wide range of speakers, expressions and character voices. The current work aims to use audiobook data to build a TTS system which produces expressive synthetic speech for different speakers.

Manually adding expression labels to the audiobook data is expensive and has typically poor inter-annotator agreement due to the high diversity of the data. This makes the SEF method difficult to use for the audiobook data. To solve this problem, automatic expression clustering methods based on acoustic features have been proposed [5, 14]. However, usually the acoustic features used for clustering contain not only expression information, but also speaker information. Therefore the expression clustering results may be influenced by speaker factors for multi-speaker data. Furthermore, the acoustic feature based clustering methods are based on the distance between acoustic features, and they are inconsistent with the training criteria used in statistical parametric speech synthesis methods, e.g. the maximum likelihood (ML) criterion.

This work presents an unsupervised SEF method. Based on the maximum likelihood (ML) criterion, the model parameter estimation and automatic expression clustering process are integrated into a single process. The speaker information is explicitly isolated from the expression clustering process. Meanwhile, the likelihood of the training data was maximized by the proposed method.

Based on the proposed unsupervised SEF method, the speaker independent expression information was extracted from multi-speaker audiobook data. It can then be combined with different speaker information to achieve expressive ebook synthesis with different speaker’s voices. For new speakers, only a small amount of neutral speech is needed to perform the speaker adaptation, while the expression information extracted from the adaptation data can be shared by any new speaker.

2. Unsupervised SEF based on CAT

The CAT model [12], has been successfully used in both speaker modelling and expression modelling.
It consists of a set of cluster models, each of which contains a set of Gaussian mean parameters while the Gaussian variances are shared over all clusters. When this CAT model is used to calculate the likelihood of an observation vector, the mean vector to be used is a linear interpolation of all the cluster means, i.e.,

$$p(o_t|\Lambda, M^{(m)}, \Sigma^{(m)}) = \mathcal{N}(o_t; \mu^{(m)}, \Sigma^{(m)})$$  \hspace{1cm} (1)$$

where $M^{(m)}$ is the matrix of $P$ cluster mean vectors for component $m$, i.e. $M^{(m)} = [\mu^{(m,1)} \ldots \mu^{(m,p)}]$ and $\Lambda$ is the CAT weight vector.

As in the standard CAT approaches the first cluster is specified as a bias cluster, thus

$$\hat{\Lambda} = [1 \; \lambda_2 \; \ldots \; \lambda_P]^{T}$$  \hspace{1cm} (2)$$

When the CAT model is used for speaker modelling, each speaker dependent information is stored in a speaker dependent CAT weight vector which forms a point in a speaker space. Similarly, for expression modelling, each expression is associated to a point in an expression space. In the case of SEF, the CAT weight vector contains both speaker and expression information. That means, some dimensions of the CAT weight vector are used to model the speakers while the others are used to model the expressions. Thus equation 1 can be re-written as

$$p(o_t|\Lambda_s, M^{(m)}_s, \Sigma^{(m)}_s) = \mathcal{N}(o_t; \mu^{(m)} + M^{(m)}_s \Lambda_s, \Sigma^{(m)}_s)$$  \hspace{1cm} (3)$$

where $\Lambda_s$ and $\Lambda_e$ are the CAT weight vectors to model the expression and speaker respectively, and $M^{(m)}_s$ and $M^{(m)}_e$ are the cluster mean matrices for component $m$ which are associated to the expression CAT weight vector and speaker CAT weight vector respectively.

2.1. Parameter estimation for unsupervised factorization

The parameter estimation of SEF can be described as estimating the independent speaker and expression parameters from the training speech in which the speaker and expression information are mixed together. When expression parameters are estimated, the speaker parameters are assumed to be known and fixed, and vice versa.

The parameter estimation algorithm for CAT based SEF was presented in [11]. However, it is assumed that the speaker and expression labels have been added for each training utterance. For highly diverse data such as audiobooks, the speaker information is usually known, but the expressions are often unknown. In this case, supervised SEF can not be used directly. In order to perform a supervised SEF with unlabelled data, an expression clustering process must be performed before the SEF process.

Automatic expression clustering methods based on acoustic features have been widely used in audiobook data processing e.g., [5, 14]. These methods group the training speech utterances into a set of clusters $\mathcal{E} = \{e_1, e_2, \ldots, e_k\}$, the speech data in each cluster is assumed to contain similar expressions. Given the expression clustering results $\mathcal{E}$ and the known speaker information, the supervised SEF can be used to estimate the expression CAT weight vector for each expression cluster by the ML criterion:

$$\hat{\Lambda}_e(\mathcal{E}) = \arg \max_{\Lambda_e(\mathcal{E})} p(O|\mathcal{H}, \mathcal{E}, \mathcal{M}, \hat{\Lambda}_e(\mathcal{E}))$$  \hspace{1cm} (4)$$

where $O$, $\mathcal{H}$ and $\mathcal{M}$ are the observation vectors, transcripts of training speech and cluster model parameters respectively, $\Lambda_s$ represents the speaker CAT weight vectors which is known and fixed, $\hat{\Lambda}_e(\mathcal{E}) = (\lambda_{e_1}^{(m)}, \lambda_{e_2}^{(m)}, \ldots, \lambda_{e_k}^{(m)})$ represents the expression CAT weight vectors based on the expression clustering results $\mathcal{E}$.

This method has two weak aspects. First, the acoustic features e.g. the mean of $F_0$ etc., are highly dependent on speakers, thus it is difficult to get the speaker independent expression clusters. Second, the expression clustering is usually based on the distance measure in the acoustic feature space, e.g. the minimum within class error, while the SEF is based on the ML criterion; thus there is an inconsistency between the parameter estimation of the two processes.

To address the problems mentioned above, this work presents a factorization method for data without expression labelling, i.e. unsupervised SEF. The proposed method integrates the expression clustering and SEF as a single process. It means that the expression clustering and the expression dependent parameter estimation are conducted simultaneously, i.e.

$$\hat{\mathcal{E}}, \hat{\Lambda}_e(\hat{\mathcal{E}}) = \arg \max_{\mathcal{E}, \Lambda_e(\mathcal{E})} p(O|\mathcal{H}, \mathcal{E}, \mathcal{M}, \Lambda_s, \Lambda_e(\mathcal{E}))$$  \hspace{1cm} (5)$$

In the CAT framework, the auxiliary function for the CAT weight estimation can be expressed as

$$Q(\hat{\Lambda}_e; \Lambda, \Lambda_s) = \sum_i (\lambda_{e_i}^{(m)} - \frac{1}{2} \lambda_{e_i}^{(m)} X_{i}^{(m)} X_{i}^{(m)T}) + C$$  \hspace{1cm} (6)$$

where $i$ is the utterance index, $C$ represents the terms independent to $\hat{\lambda}$ and the sufficient statistics $X^{(i)}$ and $y^{(i)}$ are given by

$$X^{(i)} = \sum_{m \in T_i} \gamma_{i}^{(m)} M^{(m)} \Sigma^{(m)1} M^{(m)}$$  \hspace{1cm} (7)$$

$$y^{(i)} = \sum_{m} (M^{(m)} \Sigma^{(m)-1} \sum_{t \in T_i} \gamma_{t}^{(m)} (o_{t} - \mu_{(m,1)}^{(m)})$$  \hspace{1cm} (8)$$

where $\gamma_{i}^{(m)}$ is the occupancy probability of component $m$ in time $t$, $\mu_{(m,1)}^{(m)}$ is the mean vector of component $m$ from the bias cluster.

In the case of SEF, to calculate the new expression CAT weight vectors $\hat{\Lambda}_e$, given the old expression CAT weight vectors $\Lambda_e$ and the fixed speaker CAT weight vectors $\Lambda_s$, the equation 6 can be re-written as

$$Q(\hat{\Lambda}_e; \Lambda_s, \Lambda_e) = \sum_i (\lambda_{s_i}^{(m)} \lambda_{e_i}^{(m)}) X_{i}^{(s)} X_{i}^{(e)T} \lambda_{e_i}^{(m)} + C$$  \hspace{1cm} (9)$$

$$\frac{1}{2} \lambda_{s_i}^{(m)} \lambda_{e_i}^{(m)} X_{i}^{(s)} X_{i}^{(e)T} \lambda_{e_i}^{(m)} + D$$

where $\lambda_{s_i}^{(m)}$ and $\lambda_{e_i}^{(m)}$ represent the speaker and expression CAT weight vectors for utterance $i$ respectively, $D$ represents the terms independent to $\lambda_{e_i}^{(m)}$. The sufficient statistics are given
by

\[ X^{(i)}_{EE} = \sum_{m,t \in T_i} \gamma^{(m)} \Sigma^{(m)-1} M^{(m)} \]

\[ X^{(i)}_{ES} = \sum_{m,t \in T_i} \gamma^{(m)} \Sigma^{(m)-1} M^{(m)} \]

\[ y^{(i)} = \sum M^{(m)} \Sigma^{(m)-1} \sum_{t \in T_i} (\alpha_t - \mu^{(m,1)}) \]  

(10)

For a particular partition of training utterances \( \mathcal{E} = \{ e_1, e_2, \cdots, e_k \} \), all the utterances in the same cluster share the expression CAT weight vectors, thus equation 9 can be expressed as

\[ Q(\Lambda_E, \alpha_E, \Lambda_S) = \sum_{j=1}^{k} \sum_{i \in e_j} (\lambda_E^{(e_j)} (y^{(i)} - X^{(i)} - X^{(i)} \lambda_E^{(i)}) - \frac{1}{2} \frac{\lambda_E^{(e_j)} X_E^{(i)} \lambda_E^{(e_j)}}{\Sigma_E^{(i)}}) \]  

(11)

The task of unsupervised SEF is to find a partition of the training data \( \mathcal{E} \) and the expression specific CAT weight vectors associated to this partition \( \alpha_E(\mathcal{E}) \) so that the value of equation 11 is maximized. This was realized by a K-means style clustering.

In the assignment step, for each training utterance \( O^i \), an expression cluster \( e(O^i) \) was assigned to it by

\[ e(O^i) = \arg \max_{e_j} \lambda_E^{(e_j)} (y^{(i)} - X^{(i)} - X^{(i)} \lambda_E^{(i)}) - \frac{1}{2} \frac{\lambda_E^{(e_j)} X_E^{(i)} \lambda_E^{(e_j)}}{\Sigma_E^{(i)}} \]  

(12)

In the update step, the expression CAT weight vector for each expression cluster was re-calculated. Differentiating the auxiliary function with respect to \( \lambda_E^{(e_j)} \) and equating to zero yields,

\[ \hat{\lambda}_E^{(e_j)} = \left( \sum_{i \in e_j} X^{(i)} \right)^{-1} \sum_{i \in e_j} (y^{(i)} - X^{(i)} \lambda_E^{(i)}) \]  

(13)

The assignment step and the update step were performed iteratively until convergence.

The proposed method can alleviate the problems in the framework of acoustic feature based expression clustering plus supervised SEF. In the proposed method, the expression clustering is based on the auxiliary function of SEF in which the speaker factor is explicitly removed; thus the speaker independent expression clustering can be achieved. At the same time, the expression clustering and CAT weight vector estimation are integrated into a single process based on the ML criterion and there is no inconsistency in the training process.

3. Experimental Results

The experiments presented here are based on publicly available audiobooks from Librivox.org. The length of the training data was about 28 hours from 4 audiobooks (2 male and 2 female speakers). The lightly supervised sentence alignment and selection method [13] was used to transform the audiobooks into usable training data. The data was further segmented into 3 types of speech units or utterances: narration, carrier and direct speech [5]. A rule based neutral data selection was performed based on acoustic features such as f0-range, RMS-amplitude-range, etc [15]. This resulted in 5 hours of neutral training data which was used to initialize the speaker clusters and the speaker CAT weight vectors. The sampling rate of the training speech was 16kHz and acoustic features consisted of 40 mel-cepstral coefficients, logF0, 21 (approximately bark scaled) BAP plus their delta and delta-delta information. The models were 5 state left-to-right multi-space probability distribution hidden semi-Markov models.

The CAT model used in this work consisted of 8 cluster models: 1 bias cluster model, 4 non-bias cluster models for speaker modelling and 3 non-bias cluster models for expression modelling. The CAT training process based on unsupervised SEF is summarised below:

1. Construct speaker cluster models using the selected neutral speech, iteratively updating the speaker decision trees, speaker CAT weight vectors and speaker cluster models until convergence.
2. Using acoustic feature based expression clustering, group the training speech into \( P_k \) clusters, where \( P_k \) is the dimension of expression CAT weight vectors.
3. Keep the speaker CAT weight vectors fixed. For each expression cluster, set CAT weight to one for that cluster and zero otherwise
4. Construct the decision tree for each expression cluster.
5. For each discrete expression state, re-estimate the expression CAT weight vector, based on equation 4.
6. Update cluster model parameters for all the clusters.
7. Goto 4 until convergence.
8. For each utterance \( i \), accumulate the expression statistics \( X^{(i)}_E, X^{(i)}_E \) and \( y^{(i)}_E \) as equation 10.
10. Re-construct the decision tree for each expression cluster.
11. Re-estimate the expression CAT weight vector, with the fixed speaker CAT weight vectors.
12. Update cluster model parameters for all the clusters.

In the training process described above, step 2-7 performed a process of expression clustering plus supervised SEF. This process constructed an initial expression space and sufficient statistics for unsupervised SEF were accumulated based on this initial expression space in step 8. Finally, the unsupervised SEF training was performed in steps 9 to 13. After the expression CAT weights and cluster models were trained, the speaker CAT weight vectors and cluster models can be re-estimated in similar fashion. However, in this work, the re-estimation of the speaker part was skipped due to limited time for computing.

In this work, the supervised adaptation was performed in the synthesis stage, as shown in figure 1. The expression information for synthetic speech was extracted from the expression adaptation data which was naturally expressive speech. To simplify the process, in this work the expression adaptation data was from one of the training speakers which was labelled as “speaker 1” in figure 1, thus the speaker CAT weights were known. Given the speaker CAT weights, the expression adaptation data was projected to a point in the expression space which associated to an expression CAT weight vector to maximize the likelihood of the expression adaptation data. Then, the generated expression CAT weights were shared over different speakers. As shown in figure 1, for a new speaker “speaker 2”, which
only had neutral speech, the speaker adaptation was performed to find a point in the speaker space which maximizes the likelihood of the speaker adaptation data. Finally, the expression CAT weights were concatenated with speaker CAT weights to generate the synthetic speech for “speaker 2” with the same expression as the data from “speaker 1”.

The first experiment investigated the performance of the proposed unsupervised SEF. Based on the unsupervised SEF and the cascade of the expression clustering plus supervised SEF, the CAT models were trained separately. Then using the supervised adaptation process in figure 1, expressive speech with different speakers was generated using two CAT models separately. The synthetic speech was generated for 6 speakers, 4 training speakers and 2 new speakers. For the training speakers, the speaker adaptation part in figure 1 can be skipped since the speaker CAT weight vectors were known. For the new speakers, the speaker CAT weight vectors were trained by 100 neutral utterances from each of them. An ABX test was performed to evaluate the synthetic speech. The naturally expressive speech used for expression adaptation was used as reference in an ABX test. The subjects listened to the synthetic speech from 2 systems and were asked which one is expressively closer to the reference speech. The ABX test set includes 75 evaluation utterances and the results for the training speakers and the new speakers are shown in table 1 and table 2 respectively.

Table 1: ABX test for unsupervised SEF, training speaker

<table>
<thead>
<tr>
<th>Spkr</th>
<th>Uns.</th>
<th>Clst.+sup.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>55.0%</td>
<td>45.0%</td>
<td>0.007</td>
</tr>
<tr>
<td>2</td>
<td>51.6%</td>
<td>48.4%</td>
<td>0.241</td>
</tr>
<tr>
<td>3</td>
<td>56.7%</td>
<td>43.3%</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>4</td>
<td>52.1%</td>
<td>47.9%</td>
<td>0.168</td>
</tr>
<tr>
<td>overall</td>
<td>54.3%</td>
<td>45.7%</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

Table 2: ABX test for unsupervised SEF, new speaker

<table>
<thead>
<tr>
<th>Spkr</th>
<th>Uns.</th>
<th>Clst.+sup.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>54.1%</td>
<td>45.9%</td>
<td>0.029</td>
</tr>
<tr>
<td>6</td>
<td>54.8%</td>
<td>45.2%</td>
<td>0.014</td>
</tr>
<tr>
<td>overall</td>
<td>54.0%</td>
<td>46.0%</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Table 1 and table 2 indicate that the proposed unsupervised SEF method achieves significantly better results than the framework of expression clustering plus supervised SEF for the expressiveness of synthetic speech.

Since the purpose of this work is expressive ebook synthesis using multiple speakers, experiments using full paragraphs were carried out based on the proposed unsupervised SEF. The contrast system was the neutral TTS system without expression modelling. It was implemented by skipping the expression adaptation part in figure 1, while only keeping the speaker adaptation part. A preference test was performed to evaluate the paragraph synthesis based on 15 paragraphs with an average of 3 utterances per paragraph. The results for training speakers and test speakers are shown in table 3 and table 4 respectively.

Table 3: Preference test for paragraph reading, training speaker

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Uns.</th>
<th>Clst.+sup.</th>
<th>Neutral</th>
<th>No pref</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>47.4%</td>
<td>28.5%</td>
<td>23.7%</td>
<td>0.007</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>45.9%</td>
<td>42.9%</td>
<td>11.2%</td>
<td>0.351</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>57.2%</td>
<td>41.6%</td>
<td>1.2%</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>43.6%</td>
<td>42.9%</td>
<td>13.5%</td>
<td>0.469</td>
<td></td>
</tr>
<tr>
<td>overall</td>
<td>50.1%</td>
<td>38.4%</td>
<td>11.6%</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Preference test for paragraph reading, new speaker

<table>
<thead>
<tr>
<th>Speaker</th>
<th>Uns.</th>
<th>Clst.+sup.</th>
<th>Neutral</th>
<th>No pref</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>61.7%</td>
<td>35.3%</td>
<td>3.0%</td>
<td>&lt;0.001</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>57.3%</td>
<td>36.7%</td>
<td>6.0%</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>overall</td>
<td>39.0%</td>
<td>55.2%</td>
<td>5.3%</td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

In the paragraph evaluation, the synthetic speech based on the proposed unsupervised SEF method achieved significantly better overall results than that from the neutral speech synthesizer. This indicates that the proposed method can improve the expressiveness of the ebook reading for multiple speakers. However, for different speakers, the improvement from the proposed method is inconsistent. A possible reason is, that the speaker part in the SEF model which was assumed to be free from expressive speech was trained on audiobook data based on the automatic neutral selection. However, this automatically selected data still included some undetected expressiveness. This may influence the performance of the SEF. For some speakers, if the neutral data contains more expressions, the final SEF results may be unstable. Therefore in future work, more strict neutral selection may be applied for speaker modelling.

4. Conclusion

This work presented an unsupervised SEF method for factorizing and modelling the independent speaker and expression information in unlabelled audiobook data. The expression clustering in the SEF was integrated in a single process based on the ML criterion. It significantly outperformed the cascade system of expression clustering and supervised SEF in the expressiveness of synthetic speech of multiple speakers. Based on the proposed method, the paragraph reading evaluation was carried out and for both training speakers and new speakers, the proposed method generated more expressive speech than a baseline system without expression modelling.
5. References


