HMM-based Speaker Interpolation

Susanne Rexeis[*] *Matthias Straka*[†]

Abstract

Speech style modeling deals with generating speech with different speaking styles or emotions. This paper describes approaches for style modeling and interpolation of speaking styles in HMM based speech synthesis. Evaluations for all described approaches are presented.

1 Introduction

Based on speech synthesis with HMMs this paper describes methods to model different emotional expressions and speaking styles and the interpolation between them. The goal is to make synthesized speech sound more natural especially different emotions.

In contrast to other ideas that base on variation of pitch, loudness and speed the methods in this paper work with context based decision trees.

Speaker interpolation generates synthesized voices that are a mix of styles (e.g. two emotional styles or two speakers with different accent or gender). Three different interpolation methods are introduced and compared.

2 Style Modeling

Yamagishi et al. [3] published a paper that compares two approaches for modeling emotional expressions and speaking styles: style-dependent and style-mixed modeling. In the style-dependent approach a model for each style is trained and the styles are linked afterward. In contrast, in style-mixed modeling only one model is trained for all styles.

Both models are generated automatically by using tree-based context clustering using a minimum description length splitting (MDL) criterion. Figure 1 shows the resulting tree for style-dependent modeling and figure 2 the one for style-mixed modeling. The following section will describe this method in more detail.

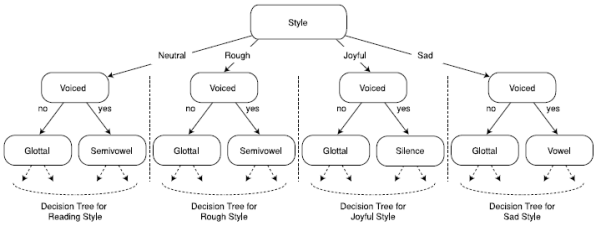

Fig. 1: decision tree for style dependent modeling

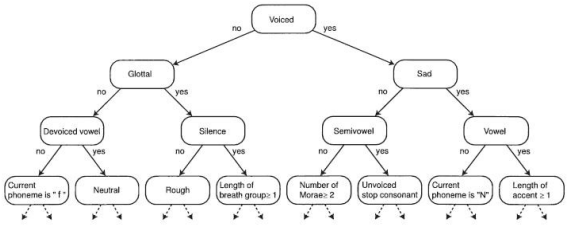

Fig. 2: decision tree for style morphing

2.1 Tree based context clustering

The trained HMM models for the phonemes are used for clustering. Both the spectrum and F_0 part of the models are clustered separately. The benefit of clustering the phoneme models is that the number of distributions needed for synthesis is strongly reduced.

The trees for clustering are generated automatically with a given set of *yes/no* questions that refer to the phonetic and linguistic context of the phoneme. These questions are used as split-

[*]susanne.rexeis@student.tugraz.at, 0330275
[†]mstraka@student.tugraz.at, 0430207

ting conditions for the PDF distributions. To a node N_m the question q is assigned that minimizes

$$\delta_m(q) = D(U') - D(U) \quad (1)$$

where $D(U')$ is the description length of model U' which is generated by splitting the node N_m question q and $D(U)$ is the description length of model U.

This difference is also used to define a stopping criteria for the automated splitting process.

2.2 Evaluation

The goal of the evaluation was to determine whether a number of test subjects could recognize the synthesized speaking style.

Four different styles were evaluated: neutral, rough, joyful and sad. Speech samples of 503 phonetically balanced sentences from the ATR Japanese speech database where taken and recorded by both a male and a female speaker. As the results were quite similar, only the ones for the male speaker are presented in this paper.

First the original recorded samples were evaluated by nine test subjects. Therefore 53 randomly chosen samples of different styles were presented and had to be classified to be either neutral, rough, joyful, sad or other. The results are shown in figure 3.

Recorded	Classification (%)				
Speech	Neutral	Rough	Joyful	Sad	Other
Neutral	96.6	2.2	0.6	0.6	0.0
Rough	2.2	96.0	0.6	0.6	0.6
Joyful	1.1	1.1	97.8	0.0	0.0
Sad	0.0	0.6	0.0	99.4	0.0

Fig. 3: Classification results of recorded speech samples

For both the style-dependent and the style mixed modeling 405 recorded speech samples of each style were used to train five-state left to right HMMs. Then the tree-based context clustering was applied. In figure 4 it can be seen that the clustering drastically reduced the number of distributions.

For evaluations 8 randomly chosen sentences drawn from 53 sentences that were not included in the training data were presented to test subjects

	Style-dependent					Style-
	Neutral	Rough	Joyful	Sad	Total	mixed
Spec. F_0 Dur	27126	27053	27164	27485	108828	108828

	Style-dependent					Style-
	Neutral	Rough	Joyful	Sad	Total	mixed
Spec.	891	752	808	926	3377	2796
F_0	1316	1269	1368	1483	5436	4404
Dur.	1070	1272	1057	950	4349	3182

Fig. 4: Number of distributions before and after clustering

who had to classify the emotion dependent on the style of speech. The test results in figure 5 and figure 6 show that the samples could be classified very well.

Synthetic	Classification (%)				
Speech	Neutral	Rough	Joyful	Sad	Other
Neutral	98.3	0.6	0.0	0.0	1.1
Rough	6.9	82.3	0.0	0.0	10.8
Joyful	1.1	0.0	94.9	0.0	4.0
Sad	0.6	1.1	0.0	94.9	3.4

Fig. 5: Classification results of style-dependent modeling

Synthetic	Classification (%)				
Speech	Neutral	Rough	Joyful	Sad	Other
Neutral	98.9	0.0	0.0	0.0	1.1
Rough	2.8	89.8	0.0	1.1	6.3
Joyful	0.6	0.0	96.0	0.0	3.4
Sad	0.0	0.6	0.0	96.0	3.4

Fig. 6: Classification results of style-mixed modeling

2.3 Advantages and Disadvantages

The style-mixed model is capable to reduce the number of distributions further than the style-dependent model as similar distributions that belong to different styles are also clustered.

During the evaluations of the two models it turned out that the samples synthesized with the

style-mixed model were considered to sound even a bit more natural than samples generated with the style-dependent model.

However, style-mixed modeling has the great disadvantage that if a new style is added the whole model has to be re-trained. In contrast a new style can be added easily to the style-dependent model where the new model is added as new subtree to the existing model.

3 Interpolation Methods

The paper about *style interpolation and morphing*[2] presents three methods to interpolate between two or more speakers. Interpolation between different emotions is essentially the same as interpolating between different speakers. The process is based on *styles* represented as HMMs with Gaussian probability distribution functions (pdfs). Therefore interpolating styles is basically done by interpolating between Gaussian PDFs.

3.1 Style Interpolation

It is possible to synthesize speech with intermediate voice characteristics between two speakers' models [4]. So Tachibana et al. [2] did the same with speaking styles.

They used N styles $S_1, S_2, \ldots, S_N$ with the mean vectors μ_k and covariance matrices $\mathbf{U}_k$. The styles are modeled with HMMs $\lambda_1, \lambda_2, \ldots, \lambda_N$. In order to control the interpolation, they used weights $a_1, a_2, \ldots, a_N$ where $\sum_{k=1}^{N} a_k = 1$. $\tilde{\mu}$ and $\tilde{\mathbf{U}}$ denote the resulting output vector and matrix.

The three interpolation methods are as follows:

(a) Interpolation among observations:

$$\tilde{\mu} = \sum_{k=1}^{N} a_k \mu_k \quad (2)$$

$$\tilde{\mathbf{U}} = \sum_{k=1}^{N} {a_k}^2 \mathbf{U}_k \quad (3)$$

(b) Interpolation among output distributions:

$$\tilde{\mu} = \sum_{k=1}^{N} a_k \mu_k \quad (4)$$

$$\tilde{\mathbf{U}} = \sum_{k=1}^{N} a_k \left(\mathbf{U}_k + \mu_k {\mu_k}^T\right) - \tilde{\mu}\tilde{\mu}^T \quad (5)$$

(c) Interpolation based on Kullback information measure:

$$\tilde{\mu} = \left(\sum_{k=1}^{N} a_k {\mathbf{U}_k}^{-1}\right)^{-1} \left(\sum_{k=1}^{N} a_k {\mathbf{U}_k}^{-1} \mu_k\right) \quad (6)$$

$$\tilde{\mathbf{U}} = \left(\sum_{k=1}^{N} a_k {\mathbf{U}_k}^{-1}\right)^{-1} \quad (7)$$

The Kullback information measure is defined as the distance between the interpolated speaking style S and each individual style S_k. It is measured between λ and λ_k:

$$I(\lambda, \lambda_k) = E_O \left[P(O|\lambda) \log \frac{P(O|\lambda)}{P(O|\lambda_k)} \right] \quad (8)$$

Using this equation we need to minimize the cost function

$$\epsilon = \sum_{k=1}^{N} a_k I(\lambda, \lambda_k) \quad (9)$$

with respect μ and $\mathbf{U}$ to end up in equations (6) and (7). Frankly speaking, these equations minimize the distance between two or more speaking styles.

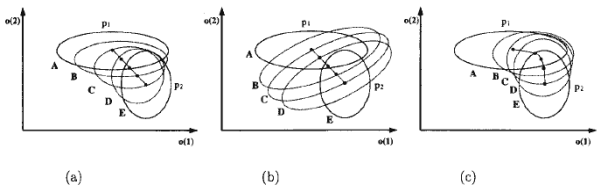

Fig. 7: Comparison between method (a), (b) and (c) with regard to interpolation between two Gaussian distributions

Figure 7 gives an idea of how the mean values and covariance matrices of the distributions change when the coefficients (a_1, a_2) are changed. From A to E they gradually change from $(1, 0)$ to $(0, 1)$.

These interpolation methods can be used to calculate the interpolated model $\tilde{\lambda}$. If all models λ_k have common structure, it is possible to calculate $\tilde{\lambda}$ directly from the λ_ks. In general, context clustering is done independently for each style, thus resulting in different structures.

To solve this problem, a text is first transformed into context-dependent phoneme labels in the synthesis stage. Then for each style, sentence HMMs

with identical topologies are created. From these the PDF sequences for spectrum, F_0 and state duration are determined. These parameters are then interpolated to obtain the desired style $\tilde{S}$. See figure 8 to get a clearer idea of this process.

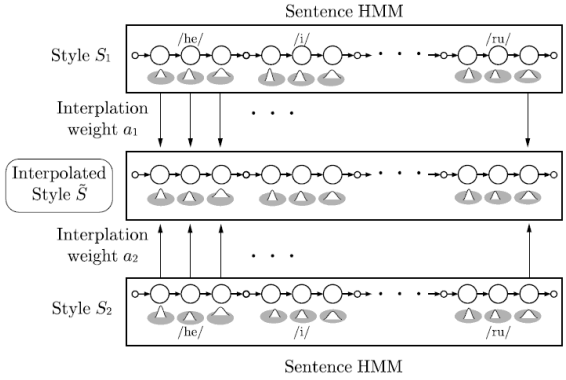

Fig. 8: Example of interpolation of two style models[2].

3.2 Style Morphing

Morphing between two styles is simply done by varying the parameters (a_1, a_2). Starting with $(1, 0)$ the parameters need to be gradually changed to $(0, 1)$ while maintaining the constraint $a_1 + a_2 = 1$. The resulting speech gradually changes as well.

3.3 Experiments

Only a few databases are available that are suitable for style interpolation. In [2] they used a database consisting of four styles: *neutral, joyful, sad* and *rough*. The database consisted of 503 sentences read by a male and a female narrator. The emotions were not real but simulated by the readers.

In the experiment they used 42 phonemes for the model training and took many phonetic and linguistic contexts into account. These included

- the number of morae in a sentence
- position of breath groups
- position of accents
- preceding, current and succeeding phonemes
- and many more

The speech signal was analyzed with the mel-cepstral analysis[1]. 25 coefficients were extracted for each 25ms window. The styles were modeled with hidden semi-Markov models with 5 left-to-right states.

In order to evaluate their results, they generated different interpolations between two styles (e.g. 50% neutral and 50% joyful) and let a group of 8 people decide what they perceived.

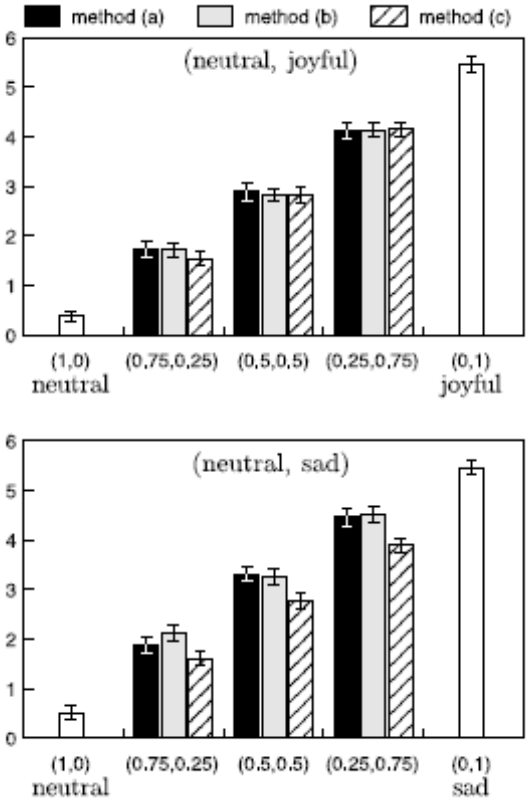

Fig. 9: Evaluation of degree of expressivity of interpolated styles[2].

In figure 9 you can see that the interpolation of two styles is really perceived as intended. While method (a) and (b) work about equally well, method (c) has a slight bias toward *neutral* speech. The bias is systematic because method (c) produces slightly faster speaking rates. The reason for this is that the calculated mean value is affected by the covariance matrices of the original distributions and

the smallest covariance becomes dominant. This can also be seen in figure 7.

4 Conclusions

In this paper we have shown how it is possible to generate decision trees that help choosing a speaking style for HMM-based speech synthesis. There are two main ways of building these trees: style-dependent and style-mixing.

Also, we have shown three ways to interpolate the generated speech between one or more styles. This technique can be used to vary the emotional context during speech or mix between genders or dialects.

For all techniques presented in this paper we provided some evaluations of their credibility. Future work may focus on different speaking styles or applications.

References

[1] T. Fukada H. Saito K. Tokuda, T. Kobayashi and S. Imai. Spectral estimation of speech based on mel-cepstral representation. *IEICE - Trans. Inf. Syst.*, J74-A:1240–1248, 1991.

[2] Makoto Tachibana, Junichi Yamagishi, Takashi Masuko, and Takao Kobayashi. Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing. *IEICE - Trans. Inf. Syst.*, E88-D(11):2484–2491, 2005.

[3] Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi. Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. *IEICE - Trans. Inf. Syst.*, E88-D(3):502–509, 2005.

[4] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Speaker interpolation for hmm-based speech synthesis system. *Journal of the Acoustical Society of Japan (E)*, 21(4):199–206, 20000700.