INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Generation of Emotion Control Vector using MDS-based Space Transformation for Expressive Speech Synthesis

Yan-You Chen[1], *Chung-Hsien Wu*[2], *Yu-Fong Huang*[2]

[1]Department of Electrical Engineering,
[2]Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan

n2896136@mail.ncku.edu.tw, chunghsienwu@gmail.com

Abstract

In control vector-based expressive speech synthesis, the emotion/style control vector defined in the categorical (CAT) emotion space is uneasy to be precisely defined by the user to synthesize the speech with the desired emotion/style. This paper applies the arousal-valence (AV) space to the multiple regression hidden semi-Markov model (MRHSMM)-based synthesis framework for expressive speech synthesis. In this study, the user can designate a specific emotion by defining the AV values in the AV space. The multidimensional scaling (MDS) method is adopted to project the AV emotion space and the categorical (CAT) emotion space onto their corresponding orthogonal coordinate systems. A transformation approach is thus proposed to transform the AV values to the emotion control vector in CAT emotion space for MRHSMM-based expressive speech synthesis. In the synthesis phase given the input text and desired emotion, with the transformed emotion control vector, the speech with the desired emotion is generated from the MRHSMMs. Experimental result shows the proposed method is helpful for the user to easily and precisely determine the desired emotion for expressive speech synthesis.

Index Terms: speech synthesis, emotion control, control vector generation

1. Introduction

Speech is the most natural way for human communication. In this sense, speech synthesis plays a critical role in speech related applications. In some situations, the speech synthesis system is required to be able to generate the speech with various expressivity. For expressive speech synthesis, the user's meaning and intention are expected to be expressed clearly by the synthetic speech with various expressions such as emotions and speaking styles. The flexibility of speech synthesis technique becomes important because of the demand for the speech with rich expressivity. Therefore, the speech synthesis based on hidden-semi Markov model (HSMM) [1]–[7] that can provide flexible speech modeling and generation is often used for the recent development of expressive speech synthesis. There are a number of related techniques proposed in the past years. Style modeling including style-dependent modeling and style-mixed modeling [8], [9] was proposed for modeling and generating certain styles with a given sufficient amount of training data. Style adaptation [10], [11] reduced the data preparation cost by using model adaptation from the neutral style to the target style. Intermediate style expressions can be also generated using style interpolation [12], [13] between two or more representative style models. Recently, style control technique [14]–[16] using multiple regression HSMM (MRHSMM) is a more sophisticated technique to control the intensity of style expressivity appearing in the synthetic speech. Conventionally, the training of MRHSMM is based on the vectors in the style space. However, no matter using style interpolation or style control methods, the weighting of the styles or the control vector of the desired emotion is difficult to be precisely defined because they are based on the categorical (CAT) emotion space as shown in Figure 1. The dimensions of the CAT emotion space are not intuitive and might be dependent upon each other.

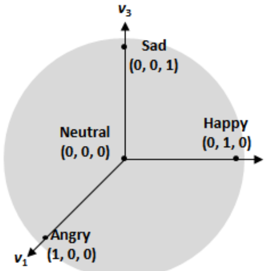

Figure 1: *The 3-dimensional CAT emotion space with four primitive emotions.*

In this sense, a more intuitive emotion space with two bipolar dimensions providing more intuitive description of emotion, which is well known as arousal-valence (AV) space [17], is taken into account in this study. Contrast to the emotions characterized in the AV space, the emotions in the CAT space, consisting of four primitive emotions, i.e. *Neutral*, *Happy*, *Angry* and *Sad*, in this study, are quite different from those in the AV space. Accordingly, it is not convincing to determine the control vector by using only the intensities of the primitive emotions in the CAT space for expressive speech synthesis. Therefore, we attempt to apply the AV emotion space to the MRHSMM-based speech synthesis framework. Using the AV space, users can easily decide their desired emotion for the synthesized speech by designating the AV values. However, the emotional characteristics for different phonemes might also be different in the CAT space. Therefore, the phoneme-based control vector should be used. To obtain the emotion control vector in the CAT emotion space in which the MRHSMM-based synthesis system is constructed, a space transformation function is then proposed to transform the designated emotion from the AV emotion space to the CAT emotion space. Before the transformation, the multidimensional scaling (MDS) [18], [19] is adopted to cope with the non-orthogonal problem by projecting the emotion representations in the AV space and

http://dx.doi.org/10.21437/Interspeech.2016-815

CAT space onto their corresponding orthogonal coordinates. Finally, the control vector in the CAT emotion space can be transformed from the defined AV values by the user for synthesizing the desired expressive speech using the MRHSMM.

The rest of this paper is organized as follows. Section II introduces the MRHSMM-based speech synthesis framework and MDS. The system overview and the proposed control vector generation are present in Section III. Section IV gives the evaluation and discussion. Conclusions are finally drawn in Section V.

2. Related Work

2.1. MRHSMM-Based Speech Synthesis

In conventional HSMM, the observation (including the static spectral and pitch features and their dynamic features) probability density functions (pdfs) $b_i(\boldsymbol{o})$ and state duration pdf $p_i(d)$ at state i are given by the mean vector $\boldsymbol{\mu}_i$ and diagonal covariance matrix $\boldsymbol{\Sigma}_i$, and mean m_i and variance σ_i^2, respectively, as follows.

$$b_i(\boldsymbol{o}) = \mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{1}$$

$$p_i(d) = \mathcal{N}(d; m_i, \sigma_i^2). \tag{2}$$

In the MRHSMM-based style control technique [14]–[16], each speech synthesis unit is modeled by the context-dependent MRHSMMs [14], in which $\boldsymbol{\mu}_i$ and m_i are respectively modeled using multiple regression as

$$\boldsymbol{\mu}_i = \boldsymbol{H}_{b_i}\boldsymbol{\xi}, \tag{3}$$

$$m_i = \boldsymbol{H}_{p_i}\boldsymbol{\xi}, \tag{4}$$

where

$$\boldsymbol{\xi} = [1, v_1, v_2, \cdots, v_L]^\top = [1, \boldsymbol{v}^\top]^\top \tag{5}$$

and $\boldsymbol{v}$ is the style vector defined in the CAT emotion space. L is the dimensionality of the style space. The component v_k of the style vector represents the degree or intensity of a certain style in speech. In addition, $\boldsymbol{H}_{b_i}$ and $\boldsymbol{H}_{p_i}$ are the regression matrices with dimension $D \times (L + 1)$ and $1 \times (L + 1)$, respectively. D is the dimensionality of $\boldsymbol{\mu}_i$.

In the training phase of the MRHSMM-based speech synthesis, the parameters of MRHSMM, i.e. $\boldsymbol{H}_{b_i}$, $\boldsymbol{\Sigma}_i$, $\boldsymbol{H}_{p_i}$, and σ_i^2, are estimated based on the least square method and the EM algorithm [15] using the training data and the corresponding style vectors as shown in Figure 1. In the speech synthesis phase, the mean parameters of each synthesis unit, $\boldsymbol{\mu}_i$ and m_i are modified based on (3) and (4) with a given desired style vector $\boldsymbol{v}$. Then the synthetic speech is generated using the HMM-based speech synthesis framework.

2.2. Multidimensional Scaling

Multidimensional scaling (MDS) [18], [19] is a set of statistical techniques used for the exploration of similarities or dissimilarities in data, and is also commonly used as a method for dimensionality reduction for large similarity or dissimilarity matrices. The goal of MDS is to find N coordinates (vectors) $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \boldsymbol{R}^Q$ s.t. the distance between the vectors best approximating the given distance matrix of N objects (i.e., vector or distribution) as follows:

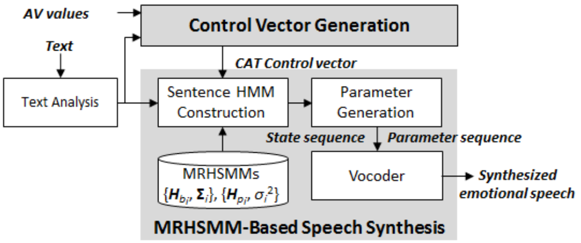
AV values
Text
Control Vector Generation
CAT Control vector
Text Analysis
Sentence HMM Construction
Parameter Generation
State sequence
Parameter sequence
MRHSMMs
$\{H_{b_i}, \Sigma_i\}, \{H_{p_i}, \sigma_i^2\}$
Vocoder
Synthesized emotional speech
MRHSMM-Based Speech Synthesis

Figure 2: *Proposed control vector generation for MRHSMM based speech synthesis*

$$\boldsymbol{\Delta} = \begin{bmatrix} \delta_{1,1} & \delta_{1,1} & \cdots & \delta_{1,N} \\ \delta_{2,1} & \delta_{2,2} & \cdots & \delta_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{N,1} & \delta_{N,2} & \cdots & \delta_{N,N} \end{bmatrix}, \tag{6}$$

where $\delta_{i,j}$, $1 \le i,j \le N$, denotes the distance between the i-th and the j-th vectors. $\delta_{n,n} = 0 \ \forall \ 1 \le n \le N$. $\boldsymbol{\Delta}$ is a symmetric matrix. In the metric MDS, a matrix $\boldsymbol{B}$ of scalar products of $\boldsymbol{\Delta}$ is defined as:

$$\boldsymbol{B} = -\frac{1}{2}\boldsymbol{H}\boldsymbol{\Delta}^2\boldsymbol{H}, \tag{7}$$

where $\boldsymbol{H} = \boldsymbol{I} - {}^1/_N\, \boldsymbol{11'}$ denotes the centralized matrix. $\boldsymbol{I}$ is an N by N identity matrix. $\boldsymbol{11'}$ is an N by N matrix of ones. By performing singular value decomposition on the matrix $\boldsymbol{B}$, the first m ($m < N$) largest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ with respect to the corresponding eigenvectors $e_1, e_2, \ldots, e_m$ can be obtained. Finally, the matrix of the projected coordinates can be obtained by $\boldsymbol{X} = \boldsymbol{E}\boldsymbol{\Lambda}^{½}$, where $\boldsymbol{E}$ and $\boldsymbol{\Lambda}$ are the matrix of m eigenvectors and the diagonal matrix of m eigenvalues of $\boldsymbol{B}$, respectively.

3. Emotion Control Vector Generation

3.1. Overview of the Proposed System

Figure 2 shows the overview of the proposed control vector generation process for the MRHSMM-based speech synthesis framework. The MRHSMMs are trained using an emotional corpus containing a large amount of speech utterances for M emotion classes along with transcripts and the corresponding control vectors in CAT emotion/style space as shown in Figure 1. First, the user inputs the text and a vector designating the AV values of the desired emotion in the AV space. Then, the context-dependent labels are analyzed by text analysis. For each phoneme in the utterance, its control vector with respect to the phone identity is transformed from the vector in the AV space to the control vector in the CAT space by using the proposed control vector generation method. According to the context dependent labels and the corresponding control vectors, the HSMM is constructed. The rest of the processes are the same as those in the standard HMM-based speech synthesis system [20]–[25]. The speech parameters are generated from the probability density function of the HSMM by parameter generation considering dynamic features [23]–[25]. Finally, the speech is generated by feeding the speech parameters through the mel log spectrum approximation (MLSA) filter-based vocoder [26], [27] followed by parameter generation. The details of the proposed control vector generation method are elaborated as follows.

3.2. Control Vector Generation

The AV space proposed by Russell [17], which is a bipolar two-dimensional space as shown in Figure 3, is more intuitive for human beings to designate the desired emotion because of its two-dimensional nature. In Figure 3, the horizontal dimension, also called valence corresponds to the emotions of pleasure/displeasure, and the vertical dimension, known as arousal, corresponds to the emotions of excitation-relaxation. Every emotion can be defined in the regions within the emotional space as a combination of valence and arousal. In this study, annotators were asked to label the AV values in the AV space according to the emotion they perceived for the emotional speech in the given emotional speech database. For each emotion, the mean vector of its distribution is chosen as its representative AV vector. However, the emotions in the emotional speech database expressed in the AV space might not be orthogonal. Therefore, the MDS is performed to obtain the vector space with m ($m = 2$ in this study) orthogonal axes (so-called MDS-AV), in which the distance matrix in the MDS procedure is based on Euclidean distance between every two representative AV vectors. The vector $\boldsymbol{y}^{(AV)}$ indicating the desired emotion in the AV space is first converted into the vector $\boldsymbol{y}^{(MDS\text{-}AV)}$ in the MDS-transformed AV space by

$$\boldsymbol{y}^{(MDS-AV)} = \boldsymbol{y}^{(AV)}\boldsymbol{T}, \tag{8}$$

where $\boldsymbol{T}$ is a regression coefficient matrix which is estimated in the calibration step, i.e.,

$$\boldsymbol{X}^{(AV)} \cdot \boldsymbol{T} = \boldsymbol{X}^{(MDS-AV)}, \tag{9}$$

where $\boldsymbol{X}^{(AV)}$ and $\boldsymbol{X}^{(MDS\text{-}AV)}$ are the matrix of the representative AV vectors in the AV space and the matrix of the corresponding vectors in the MDS-transformed AV space obtained from the emotional speech database, respectively.

With respect to acoustic features, different phonemes might have their own CAT emotion space. Therefore, for each phoneme, its MDS-transformed CAT emotion space (MDS-CAT) with m ($m = 2$) orthogonal axes is constructed based on the acoustic distances between every two categorical emotions. For each phoneme, the acoustic features of each emotion are represented by its context-free HSMM, which is the average of the HSMMs restored from the MRHSMMs according to the defined control vector and context-dependent labels for the phoneme in the training corpus. For each HSMM, each state is described by a Gaussian mixture model. Therefore, in MDS, the distances between every two HSMMs in the distance matrix are calculated by the symmetric Kullback-Leibler divergence (KLD) [19], [28] as follows

$$\delta_{i,j} = \frac{1}{Q}\sum_{q=1}^{Q} D_{KL}\left(\mathcal{N}_i^{(q)} \middle\| \mathcal{N}_j^{(q)}\right) + D_{KL}\left(\mathcal{N}_j^{(q)} \middle\| \mathcal{N}_i^{(q)}\right), \tag{10}$$

where Q denotes the number of states in HSMM. $D_{KL}(\mathcal{N}_i^{(q)} \| \mathcal{N}_j^{(q)})$ is the KLD between two distributions $\mathcal{N}_i^{(q)}$ and $\mathcal{N}_j^{(q)}$ at state q. For converting the vector from the MDS-AV into MDS-CAT, a GMM-based conversion function [29], [30] is used as follows

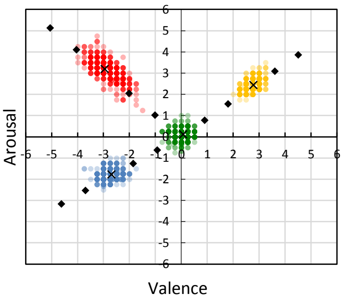

Figure 3: *Emotion distributions in the AV emotion space. The yellow, red, blue, and green distributions denote the distributions of AV values of happy, angry, sad, and neutral, respectively. The black crosses represent the mean vectors for four emotions.*

$$\begin{aligned}
&\boldsymbol{y}^{(MDS-CAT)} \\
&= \sum_{e\in\boldsymbol{E}} \frac{\mathcal{N}\left(\boldsymbol{y}^{(MDS-AV)};\boldsymbol{\mu}_e^{(MDS-AV)},\boldsymbol{\Sigma}_e^{(MDS-AV)}\right)}{\sum_{f\in\boldsymbol{E}}\mathcal{N}\left(\boldsymbol{y}^{(MDS-AV)};\boldsymbol{\mu}_f^{(MDS-AV)},\boldsymbol{\Sigma}_f^{(MDS-AV)}\right)} \\
&\times\left[\left(\boldsymbol{\Sigma}_e^{(MDS-CAT)}\right)^{1/2}\left(\boldsymbol{\Sigma}_e^{(MDS-AV)}\right)^{-1/2}\left(\boldsymbol{y}^{(MDS-AV)} - \boldsymbol{\mu}_e^{(MDS-AV)}\right)\right. \\
&\left.+\boldsymbol{\mu}_e^{(MDS-CAT)}\right],
\end{aligned} \tag{11}$$

where $\boldsymbol{E} = \{happy, angry, sad, neutral\}$ is the emotion set. $\boldsymbol{\mu}_e^{(MDS\text{-}AV)}$ and $\boldsymbol{\Sigma}_e^{(MDS\text{-}AV)}$ denote the mean vector and the covariance matrix of emotion e in the MDS-transformed AV space, respectively. $\boldsymbol{\mu}_e^{(MDS\text{-}CAT)}$ and $\boldsymbol{\Sigma}_e^{(MDS\text{-}CAT)}$ denote the mean vector and the covariance matrix of emotion e in the MDS-transformed CAT space for the current phoneme, respectively. $\boldsymbol{\Sigma}_e^{(MDS\text{-}CAT)}$ and $\boldsymbol{\Sigma}_e^{(MDS\text{-}AV)}$ are both diagonal matrices.

Finally, the control vector $\boldsymbol{v}$ in CAT emotion space is obtained from the transformed vector $\boldsymbol{y}^{(MDS\text{-}CAT)}$ by the following interpolation function:

$$\begin{aligned}
\boldsymbol{v} = \sum_{e\in\boldsymbol{E}'} & S\left(\boldsymbol{y}^{(MDS-CAT)},\boldsymbol{\mu}_e^{(MDS-CAT)}\right) \\
&\times I\left(\boldsymbol{y}^{(MDS-CAT)},\boldsymbol{\mu}_e^{(MDS-CAT)}\right)\times\boldsymbol{v}_e^{(CAT)},
\end{aligned} \tag{12}$$

where $\boldsymbol{E}' = \{e : e\in\boldsymbol{E} \text{ and } e \neq neutral\}$. $\boldsymbol{v}_e^{(CAT)}$ is the control vector of the defined emotions which are orthogonal in the CAT space for the MRHSMM. In this study, the defined emotions contain *happy*, *angry*, *sad*, and *neutral* emotions, and their control vectors are defined as (0, 1, 0), (1, 0, 0), (0, 0, 1), and (0, 0, 0), respectively. $S(\cdot, \cdot)$ is the similarity function used to evaluate the contribution of each emotion vector defined as:

$$\begin{aligned}
&S(\boldsymbol{y}^{(MDS-CAT)},\boldsymbol{\mu}_e^{(MDS-CAT)}) \\
&= \frac{1+\boldsymbol{cos}\left(\boldsymbol{y}^{(MDS-CAT)} - \boldsymbol{\mu}_{neutral}^{(MDS-CAT)},\boldsymbol{\mu}_e^{(MDS-CAT)} - \boldsymbol{\mu}_{neutral}^{(MDS-CAT)}\right)}{\sum_{f\in\boldsymbol{E}'} 1+\boldsymbol{cos}\left(\boldsymbol{y}^{(MDS-CAT)} - \boldsymbol{\mu}_{neutral}^{(MDS-CAT)},\boldsymbol{\mu}_f^{(MDS-CAT)} - \boldsymbol{\mu}_{neutral}^{(MDS-CAT)}\right)}
\end{aligned} \tag{13}$$

and $\boldsymbol{cos}$ is the cosine function. $I(\cdot, \cdot)$ is the intensity function used to evaluate the intensity of each emotion vector defined as:

$$I\left(\boldsymbol{y}^{(MDS-CAT)},\boldsymbol{\mu}_e^{(MDS-CAT)}\right) = \frac{D\left(\boldsymbol{y}^{(MDS-CAT)},\boldsymbol{\mu}_{neutral}^{(MDS-CAT)}\right)}{D\left(\boldsymbol{\mu}_e^{(MDS-CAT)},\boldsymbol{\mu}_{neutral}^{(MDS-CAT)}\right)} \tag{14}$$

and $D(\cdot, \cdot)$ is the Euclidean metric.

4. Evaluation

4.1. Speech Corpus and System Construction

For evaluating the proposed system, an emotional speech corpus, called MHMC-EMO, which contains approximately 4,000 utterances uttered by a female speaker for four emotions (i.e. *happy*, *angry*, *sad*, and *neutral*) was collected. There are approximately 1,000 utterances for each emotion. For training the MRHSMM, the CAT control vector is required as well as the speech and its corresponding full context labels. For each utterance, its control vector is provided according to its emotion class. Besides, in order to obtain the emotion distributions in the AV space, four annotators were asked to label the AV values for each utterance. The average of the four sets of the AV values was used as the result of the AV values for this utterance. The distributions of the emotions are therefore calculated and shown in Figure 3.

4.2. System Evaluation and Comparisons

4.2.1. System Friendliness Test

The mean opinion score (MOS) test was conducted to evaluate the degree of system friendliness. In the evaluation, fifty participants were invited to use these two systems for a while. Two emotion spaces were used to define the desired emotion: 1) CAT emotion space (**CAT**) as shown in Figure 1. 2) **AV**, i.e., AV space in this study, as shown in Figure 3, are included in this evaluation. For the MOS test, each participant was asked to score each system from 1 (very inconvenient to define the desired emotion) to 5 (very convenient to define the desired emotion) after using the system. The result is shown in Figure 4. The result shows that the AV space is more intuitive then the CAT emotion space for defining the desired emotion. According to the feedbacks from the participants, the desired emotion except the primitive emotions, i.e., *happy*, *angry*, *sad*, and *neutral*, can be more easily decided by using the AV space. For the primitive emotions, the results of using the AV and the CAT emotion spaces are similar.

4.2.2. Emotion Perception Test

In this evaluation, fifty participants were invited to score the systems by using MOS according to how close the emotions they perceived from the generated speech as they expected. Ten speech utterances were synthesized by each system and evaluated by each participant. Three systems including 1) **CAT**: the system using the CAT emotion space, 2) **AV**: the system using the AV space without MDS, i.e., it use only (12) to convert the vector into the CAT emotion space, and 3) **AV+MDS**: the system using the AV space and the proposed emotion control vector generation, are included in this evaluation. The result is shown in Figure 5. The MOS result of **AV+MDS** achieved the highest score. The MOS results of **CAT** and **AV** are similar. It shows that MDS-based transformation is effective to generate the speech with the user-expected emotion.

4.2.3. Emotion Intensity Test

In the third experiment, we evaluated the speech synthesized by changing the AV value as shown in Figure 6 (also presented by the black diamonds in Figure 3). The test synthetic speech was evaluated by comparing it with the reference speech synthesized by the mean vector of the emotion in the AV space as shown in Figure 3. The participants were asked to evaluate the test speech in a five point scale, where 5 for the speech with specified emotion is well expressed, 3 for the speech with almost the same emotion as the reference, and 1 for the speech similar to the neutral emotion. We calculate each score by

$$Score = \frac{\sum_{s=1}^{5} s \times (Number\ of\ times\ that\ s\ was\ chosen)}{Number\ of\ evaluation\ pairs}, \quad (15)$$

where s is the evaluated score by the participants. The result in Figure 6 shows the degree of the emotion that can be controlled by changing the corresponding AV values.

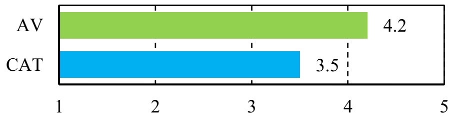

Figure 4: *MOS results of the system friendliness test.*

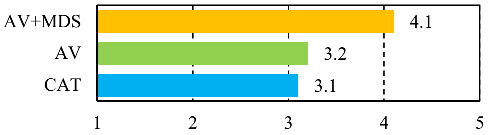

Figure 5: *MOS results of emotion perception test.*

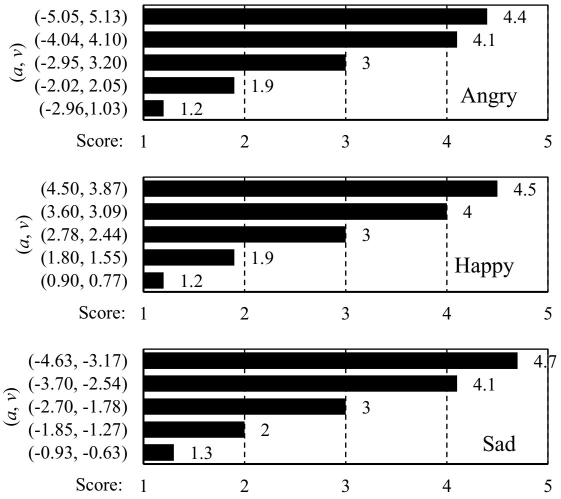

Figure 6: *Results of emotion intensity test.*

5. Conclusions

In this paper, we proposed a method for helping user to designate his/her desired emotion for expressive speech synthesis. In this method, both the AV space and the CAT space are considered. Besides, MDS is adopted to project the AV space and the CAT emotion space to their corresponding orthogonal coordinates for space transformation. An approach to emotion control vector generation is proposed for transforming the control vector in the AV space to the control vector in the CAT space for MRHSMM-based expressive speech synthesis. From the experimental results, the AV space is more intuitive for user to designate the desired emotion for synthesizing various emotions. Compared to the baseline MRHSMM-based system and the system using only AV space, the system with the proposed MDS-based control vector generation is more promising to generate the speech with various emotions.

6. References

[1] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system version 2.0," in *Proc. ISCA SSW6*, Bonn, Germany, Aug. 2007, pp. 294–299.

[2] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] M. Schröder, "Emotional speech synthesis: A review," in *Proc. EUROSPEECH 2001*, Aalborg, Denmark, Sept. 2001, pp. 561–564.

[4] T. Nose and T. Kobayashi, "Recent development of HMM-based expressive speech synthesis and its applications," in *Proc. APSIPA ASC 2011*, PID:189, Xi'an, China, Nov. 2011.

[5] J. Yamagishi, T. Masuko, and T. Kobayashi, "HMM-based expressive speech synthesis—towards TTS with arbitrary speaking styles and emotions," in *Proc. SWIM 2004*, Maui, USA, Jan. 2004.

[6] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 14, no. 4, pp. 1109–1116, July 2006.

[7] C.-H. Wu, C.-C. Hsia, C.-H. Lee, and M.-C. Lin, "Hierarchical prosody conversion using regression-based clustering for emotional speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6 pp. 1394–1405, August 2010.

[8] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sept. 1-4, 2003, pp.2461–2464.

[9] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 3, pp. 503–509, 2005.

[10] J. Yamagishi, M. Tachibana, T. Masuko, and T. Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," in *Proc. ICASSP 2004*, Montreal, Quebec, Canada, May 17-21, 2004, vol. 1, pp. I-5-8.

[11] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE Trans. Inf. & Syst.*, vol. E89-D, no. 3, pp. 1092–1099, Mar. 2006.

[12] M. Tachibana, J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "HMM-based speech synthesis with various speaking styles using model interpolation", in *Proc. Int. Conf. Speech Prosody*, Nara, Japan, Mar. 23-26, 2004, pp.41–3–416.

[13] M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 11, pp. 2484–2491, Nov. 2005.

[14] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," in *Proc. ICASSP 2001*, Salt Lake City, Utah, USA, May 7-11, 2001, pp. 513–516.

[15] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 9, pp. 1406–1413, Sept. 2007.

[16] T. Nose and T. Kobayashi, "A perceptual expressivity modeling technique for speech synthesis based on multiple-regression HSMM," in *Proc. INTERSPEECH 2011*, Florence, Italy, Aug. 27-31, 2011, pp. 109-112.

[17] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[18] T. F. Cox and M. A. A. Cox, Multidimensional Scaling. Chapman Hall, London, 1994.

[19] C.-C. Hsia, K.-Y. Lee, C.-C. Chuang, and Y.-H. Chiu, "Multidimensional Scaling for Fast Speaker Clustering," in *Proc. ISCSLP 2010*, Tainan, Taiwan, Nov. 29-Dec. 3, 2010, pp. 296–299.

[20] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. EUROSPEECH 1999*, vol. 5, Budapest, Hungary, Sept. 5-9, 1999, pp. 2347–2350.

[21] C.-C. Hsia, C.-H. Wu, and J.-Y. Wu, "Exploiting prosody hierarchy and dynamic features for pitch modeling and generation in HMM-based speech synthesis." *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 1994-2003, Nov. 2010.

[22] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "A hidden semi-Markov model-based speech synthesis system," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 825–834, May 2007.

[23] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis from HMMs using dynamic features," in *Proc. ICASSP 1996*, Atlanta, Georgia, USA, May 7-10, 1996, vol. 1. pp. 389-392.

[24] K. Tokuda, T. Yoshimura, T. Masuko, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, Istanbul, Turkey, Jun. 5-9, 2000, pp. 1315–1318.

[25] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. & Syst.*, vol. E90-D, no. 5, pp. 816–824, May 2007.

[26] T. Kobayashi, S. Imai, and T. Fukuda, "Mel-generalized log spectral approximation filter," *IEICE Trans. Fund.*, vol. J68-A, no. 6, pp. 610-611, 1985.

[27] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," in *Proc. ICASSP 1992*, San Francisco, CA, USA, Mar. 23-26, 1992, vol. 1, pp. 137–140.

[28] J. Goldberger and H. Aronowitz, "A distance measure between GMMs based on the unsented transform and its application to speaker recognition," in *Proc. EUROSPEECH 2005*, Lisboa, Portugal, Sept. 4-8, 2005, pp. 1985-1988.

[29] T. Toda, A. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[30] C.-C. Hsia, C.-H. Wu, and J.-Q. Wu, "Conversion Function Clustering and Selection Using Linguistic and Spectral Information for Emotional Voice Conversion," *IEEE Trans. Comp.*, vol. 56, no. 9, pp. 1245-1253, Sept. 2007.