EMOTIONAL SPEECH SYNTHESIS USING SUBSPACE CONSTRAINTS IN PROSODY

Shinya Mori, Tsuyoshi Moriyama, Shinji Ozawa

Keio University
Department of Information and Computer Science,
3–14–1 Hiyoshi, Yokohama-shi, Kanagawa 223–8522, Japan
kpeg@ozawa.ics.keio.ac.jp

ABSTRACT

An efficient speech synthesis method that uses subspace constraint in prosody is proposed. Conventional unit selection methods concatenate speech segments stored in database, that require enormous number of waveforms in synthesizing various emotional expressions with arbitrary texts. The proposed method employs principal component analysis to reduce the dimensionality of prosodic components, that also allows us to generate *new* speech that are similar to training samples. The subspace constraint assures that the prosody of the synthesized speech including F0, power, and speech length hold their correlative relation that training samples of emotional speech have. We assume that the combination of the number of syllables and the accent type determines the correlative dynamics of prosody, for each of which we individually construct the subspace. The subspace is then linearly related to emotions by multiple regression analysis that are obtained by subjective evaluation for the training samples. Experimental results demonstrated that only 4 dimensions were sufficient for representing the prosodic changes due to emotion at over 90% of the total variance. Synthesized emotion were successfully recognized by the listeners of the synthesized speech, especially for "anger", "surprise", "disgust", 'sorrow", "boredom", "depression", and "joy".

1. INTRODUCTION

Text-to-speech technology has reached the state that to consider its human-competitive quality to be realized in the near future [1]. The quality refers to rich expression with emotion that also retains the speaker's personality, voice quality, and naturalness. The application of such a technology ranges over humanoid helper robots, live video conference, and friendly human-machine interface.

Past studies have worked on parametric models for synthesis but they have been suffered from the synthetic quality due to the limitation of the models. Unit selection methods [2] instead have been on the rise due to its high quality. A drawback of this approach is that it is only able to synthesize what are stored in the database. The more variations to synthesize, the more wave forms need to be collected [3]. Automatic generation of the concatinated units according to the context [4] is not always effective for non-categorical context such as emotion that varies.

An isolated word spoken with a variety of emotions share the same linguistic conditions, so the speech samples should be similar to each other. Considering the central limit theorem in the observation suggests us modeling the speech distribution by the set of mean and variance, that is the model assumed in principal component analysis (PCA). Past studies have used PCA mainly for representing the variation in speakers [5]. Representing speech samples by the subspace spanned by the principal components has many benefits: the dimensionality of the variables in speech can be reduced because they are correlated, *new* speech that are similar to the training samples can be generated, and the subspace constraints the speech variables of the synthesized speech to hold their correlative dynamics.

An efficient speech synthesis method that uses subspace constraint in prosody is proposed. In order to focus on emotion among other contextual factors, we work on words isolatedly spoken by a single speaker. We assume that the combination of the number of syllables and the accent type determines the correlative dynamics in prosody. Given emotion to synthesize, the system generates the prosody pattern, and the waveform using TD-PSOLA [6].

2. GENERATING PROSODY FROM EMOTION

The proposed method assumes that the emotional content determines the prosody pattern of the speech. The prosody pattern is a high dimensional space whose components such as F0 and power correlate each other in conveying emotion. The proposed method reduces the dimensionality and constructs the low-dimensional subspace that holds the correlative relation between prosodic parameters at any location in the space. Assuming that the combination of the accent type and the

This study was performed through Special Coordination Funds of the Ministry of Education, Culture, Sports, Science and Technology of the Japanese Government.

number of syllables determines the subspace, the training samples contain a representative word for each combination that is spoken with a variety of emotions.

2.1. Constructing the subspace of prosody patterns

Each training sample gives the prosody pattern, including the F0 contour, the power contour, and the lengths of the syllables, denoted as $\mathbf{p}_i = \{ f_{i1}, f_{i2}, \cdots, f_{iL}, a_{i1}, a_{i2}, \cdots, a_{iL}, l_{i1}, l_{i2}, \cdots, i_{in} \}^T$ ($i = 1, ..., N$, N: the number of training samples, L: speech length). In constructing the subspace, the lengths of F0 and power contours are both normalized to L_0 (proportional to the number of syllables), and the intensities are so normalized as to have zero mean and unit variance. PCA computes the eigenvectors, $\mathbf{v}_j$, with corresponding eigenvalues, λ_j (sorted as $\lambda_j \geq \lambda_{j+1}$), that span the subspace. Denoting the mean prosody pattern of the training samples as $\bar{\mathbf{p}}$, a prosody pattern $\mathbf{p}$ is represented by

$$\mathbf{p} = \bar{\mathbf{p}} + \sum_{j=1}^{m} c_j \cdot \mathbf{v}_j, \tag{1}$$

where c_j is the j-th prosody parameter (the principal component score). The number of eigenvectors, m, is so determined as the cumulative contribution of the total variance exceeds a preset threshold.

2.2. Predicting prosody parameters from emotion

An emotion vector $\mathbf{e} = \{ e_1, e_2, \cdots, e_K \}^T$ (K: the number of emotions to synthesize) contains the degree of how much each emotion is perceived from the synthesized speech. Multiple regression analysis estimates the coefficients that linearly transform the specified emotion vector into the prosody parameters as

$$\mathbf{c} = \mathbf{R}\,\mathbf{e}, \tag{2}$$

where $\mathbf{R}$ is a set of multiple regression coefficients. The prosody pattern to be synthesized is obtained further using (1).

3. SYNTHESIZING SPEECH USING SUBSPACE

Fig. 1 illustrates the schematic overview of the proposed system. Given the word and the emotion vector to synthesize, the system selects a set of eigenvectors correspondent with the combination of the number of syllables and the accent type of the word. The system transforms the emotion vector into the prosody parameters using (2), and further into the prosody pattern to be synthesized using (1). The system compensates the power of consonants within three classes including fricative consonants such as /s/, plosive ones such as /t/, and voiced ones such as /m/. This rule has empirically been determined. Finally, TD-PSOLA synthesizes the speech from the prosody pattern using phoneme database that are obtained from neutral speech.

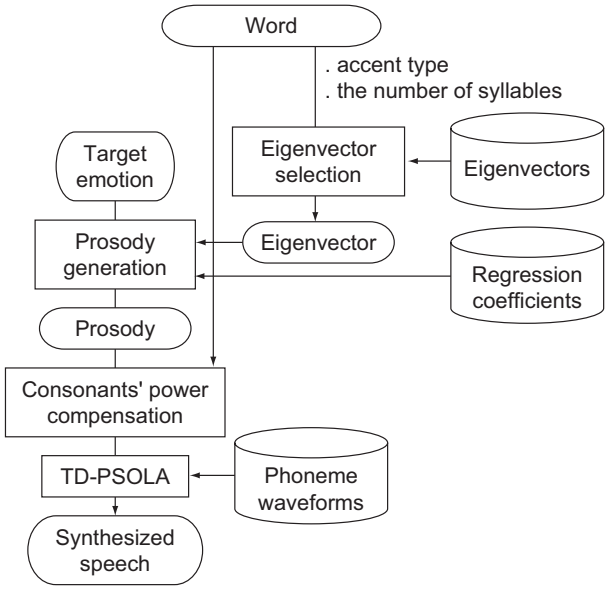


Fig. 1. The proposed speech synthesis system

4. EXPERIMENTS

We examined the subspaces constructed, the precision of the regression coefficients, and the emotional speech synthesized. Word utterances by one person were used for avoiding the effects of other contexts than emotion. For 20 combinations of the number of syllables (up to 6) and the accent types possible for each (Table 1), an actor tried to express 47 emotions listed in Table 2 (940 samples in total). Speech data were digitally recorded at 16 kHz in 16-bit levels. F0 was detected by cepstrum analysis using 32ms Hamming window at every 2ms. We empirically chose $L_0 = 100n$ (n: the number of syllables).

4.1. Evaluation of the subspaces constructed

Table 3 shows an example set of cumulative proportions of total variance for the extracted principal components. The result shows that the original 1005 dimensional space of prosody for 5 syllable speech can be represented only by 4 dimensions at 92.6% of the total variance. A set of F0 contours when the prosody parameters c_1 and c_2 in (1) change from $-3\sqrt{\lambda_1}$ to $3\sqrt{\lambda_1}$ and from $-3\sqrt{\lambda_2}$ to $3\sqrt{\lambda_2}$, respectively, are shown in Fig. 2(a) and Fig. 2(b). Both F0 and power wholly decreased and the length got elongated as c_1 increased (an *emphasizer*), while the head of F0 contour and the tail of power contour were lifted and the length got elongated as c_2 increased (a *steepener*). Other examples showed the same.

4.2. Evaluation of the regression coefficients

12 emotions that were found by cluster analysis to be approximately uncorrelated with each other (Table 4) were used in

Table 1. Words used for constructing the subspaces

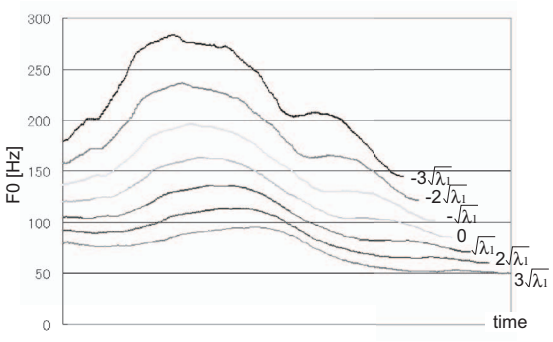
# of syllables	accent types	representatives
2	HL	/nama/
	LH	/nami/
3	HLL	/midori/
	LHL	/naname/
	LHH	/nagame/
4	HLLL	/arawani/
	LHLL	/amamizu/
	LHHL	/arayuru/
	LHHH	/omonaga/
5	HLLLL	/naniyorimo/
	LHLLL	/amamizuwa/
	LHHLL	/amanogawa/
	LHHHL	/yawarageru/
	LHHHH	/amarimono/
6	HLLLLL	/emoiwarenu/
	LHLLLL	/omoumamani/
	LHHLLL	/amagaeruwa/
	LHHHLL	/iwazumogana/
	LHHHHL	/oborozukiyo/
	LHHHHH	/warawaremono/

the subjective experiment. 20 subjects rated each one of 940 samples in seven steps as to how much these emotions were relatively perceived to the neutral. The average across the subjects were computed after removing outliers to obtain the emotion vectors.

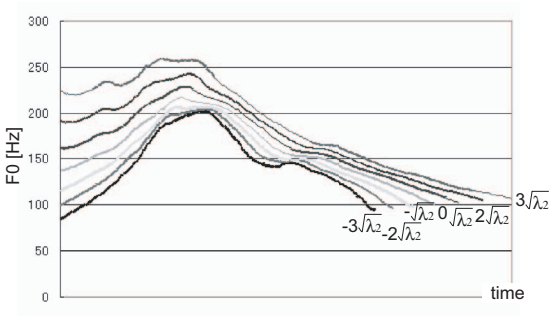
Table 3 shows an example set of coefficients of determination. Up to the second principal component that cover about 80% of the total variance in prosody (Table 3), the coefficients were larger than 0.8. It indicates that emotion vectors can predict prosody patterns of emotional speech..

4.3. Evaluation of the synthesized speech

Table 5 shows how well the synthesized speech were recognized to contain the synthesized emotion, with synthesized emotions as row labels and recognized emotions as column labels. Testing speech samples included 20 words (10 used in training data and 10 otherwise) with 12 emotions to make 240 samples in total. 20 subjects rated them and . Diagonal boxes indicate that the intended emotions were perceived from the synthesized speech. The result shows that “anger”, “surprise”, “disgust”, ‘sorrow”, “boredom”, “depression”, and “joy” were successfully recognized by the listeners. “contempt” and “funny” may need the spectrum envelop to be incorporated.



(a) c_1 changes from $-3\sqrt{\lambda_1}$ to $3\sqrt{\lambda_1}$



(b) c_2 changes from $-3\sqrt{\lambda_2}$ to $3\sqrt{\lambda_2}$

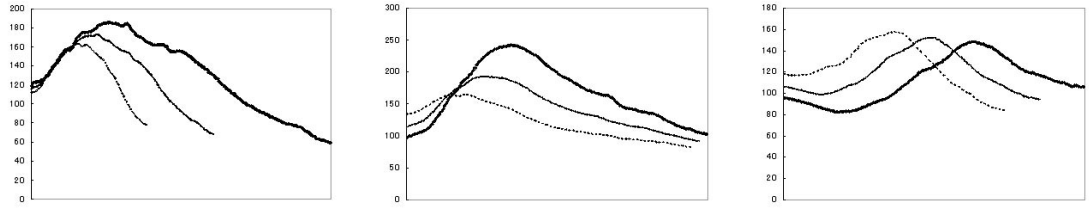
Fig. 2. F0 contours for different values of the prosody parameters.

5. CONCLUSION

An efficient speech synthesis method that uses subspace constraint in prosody was proposed. The system transforms the target emotion into prosody parameters using multiple regression equation, further into the prosody pattern using the eigenvectors of the subspace, and synthesize the wave form using TD-PSOLA. PCA dramatically reduced the dimensionality and succeeded in modeling the correlative relation between prosody components in conveying emotion. Experimental results demonstrated that the intended emotions were perceived from the synthesized speech, especially “anger”, “surprise”, “disgust”, ‘sorrow”, “boredom”, “depression”, and “joy”. Future work includes incorporating voice quality in addition to prosody, compensating the duration of phonemes, and applying the proposed framework to other context factors.

6. REFERENCES

[1] Gerard Bailly, Nick Campbell, and Bernd Mobius, “ISCA special session: Hot topics in speech synthesis,” in *Eurospeech 2003*, September 2003.



(a) /imamade/, pitch contour, "boredom", $e_{10} = 0 \rightarrow 3 \rightarrow 6$

(b) /arawani/, pitch contour, "joy", $e_2 = 0 \rightarrow 3 \rightarrow 6$

(c) /naname/, pitch contour, "disgust", $e_3 = 0 \rightarrow 3 \rightarrow 6$

Fig. 3. An example of the synthesized F0 pattern

Table 2. Emotions the speaker intended to include

neutral	envy	contempt	expectation
anger	irritated	pleased	happy
joy	complaining	cynical	fond
disgust	longing	uninterested	dislike
scorn	synpathising	admiring	reluctance
funny	tolerance	pride	depressed
worried	chuckling	love	accusation
gentle	depressed	dolor	anxiety
relief	scold	flattery	surprise
indignation	sorrow	satisfied	hasty
shameful	scared	boredom	shocked
calmness	hateful	suffering	

Table 3. An example set of cumulative proportions of the total variance, and coefficients of determination for /naniyorimo/ (5 syllables and HLLLL-accent)

principal components	cumulative proportions	coefficients of determination
λ_1	58.7%	0.837
λ_2	79.6%	0.810
λ_3	88.6%	0.554
λ_4	92.6%	0.475
$\vdots$	$\vdots$	

Table 4. Emotions used in the subjective experiment

e_1	anger	e_7	depression
e_2	joy	e_8	funny
e_3	surprise	e_9	sorrow
e_4	disgust	e_{10}	boredom
e_5	contempt	e_{11}	suffering
e_6	pride	e_{12}	shame

Table 5. An example set of subjective evaluations of the synthesized speech.

	ang.	joy	dsg.	srp.	cnt.	prd.	dpr.	fnn.	srr.	brd.	sff.	shm.
anger	**19**	0	0	0	0	1	0	0	0	0	0	0
joy	4	**11**	2	4	0	0	0	0	0	0	0	0
disgust	2	0	**13**	0	0	0	5	0	0	0	0	0
surprise	2	0	0	**18**	0	0	0	0	0	0	0	0
contempt	4	0	0	10	**4**	2	0	0	0	0	0	0
pride	4	1	1	3	1	**6**	0	0	2	1	0	0
depression	0	0	0	0	1	0	**14**	0	0	5	0	0
funny	2	2	0	7	2	5	0	**0**	1	1	0	0
sorrow	0	0	0	0	0	0	0	0	**20**	0	0	0
boredom	0	0	0	0	0	0	0	0	0	**20**	0	0
suffering	4	1	0	0	3	3	2	0	5	2	**2**	1
shame	0	4	1	0	0	1	5	1	2	5	0	**1**

[2] E. Eide, A. Aaron, R. Bakis, W. Hamza, M. Pichemy, and J. Pitrelli, "A corpus-based approach to <ahem/> expressive speech synthesis," in *5th ISCA Speech Synthesis Workshop*, June 2004.

[3] Akemi Iida, Nick Campbell, Soichiro Iga, Fumito Higuchi, and Michiaki Yasumura, "A speech synthesis system for assisting communication," in *Proceedings of the ISCA Workshop on Speech and Emotion*, September 2000, pp. 167–172.

[4] M Akamine and T Kagoshima, "Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (tos drive tts)," in *Proceedings of International Conference on Spoken Language Processing*, 1998, vol. 5.

[5] R. Kuhn, J.C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 695–707, 2000.

[6] Eric Moulines and Francis Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, no. 5-6, pp. 453–467, 1990.