EMOTIONAL SPEECH SYNTHESIS USING SUBSPACE CONSTRAINTS IN PROSODY

Shinya MORI Tsuyoshi MORIYAMA Shinji OZAWA
Dept. of Information & Computer Science, Keio University, JAPAN

Synthesis of emotional speech - past studies

context → prosody

Formant synthesis	Concatenative synthesis
pros : any emotion	pros : natural
cons : artificial	cons : only stored

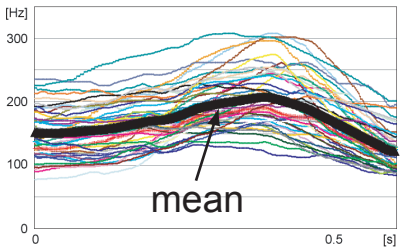
Motivation

"How can you synthesize natural speech that conveys any kinds of emotion with their gradation?"

Observation

1. mean + variance

ex.) /a/ /ra/ /yu/ /ru/



F0 contours with various emotion

2. the number of morae and the position of accent determine the variance

ex.)

/nanname/ (LHL)

vs.

/naniyorimo/ (HLLLL)

Basic idea of the proposed method

. PCA gives a statistical model for the motions in prosody
. The model is trained for each combination of the number of morae and the position of accent

Subspace constrained generation of prosody

Training phase

A male speaker tried lots of emotions (47) for each combination of the number of morae (2-6) and the position of accent

Extract prosody and project into subspace

$$\mathbf{p}_i = [f_{i1}, f_{i2}, \ldots, f_{iL}, a_{i1}, a_{i2}, \ldots, a_{iL}, l_{i1}, l_{i2}, \ldots, L_{in}], \quad (1)$$

f ... F0 a ... power l ... mora length i ... i-th training sample
L ... speech length n ... the number of morae

$$\mathbf{p}_i = \bar{\mathbf{p}} + \Sigma\, c_j * \mathbf{v}_j\,, \quad (2)$$

c_j ... j-th principal component score
$\mathbf{v}_j$ ... eigen vector of j-th principal component

Evaluate emotional content by subjective experiment

$$\mathbf{e} = [e_1, \ldots, e_K], \quad K \text{ ... the number of emotions} \quad (3)$$

Relate them

$$\mathbf{c} = \mathbf{R}\,\mathbf{e}, \quad \mathbf{R} \text{ ... partial regression coefficients} \quad (4)$$

Synthesis phase

$\mathbf{e} \xrightarrow{(4)} \mathbf{c} \xrightarrow{(2)} \mathbf{p} \xrightarrow{\text{TD-PSOLA}}$ waveform

$\mathbf{v}$ in (2) and $\mathbf{R}$ in (4) depend on the word

Results and Conclusion

. "Anger", "surprise", "disgust", "sorrow", "boredom", "depression" were synthesized well.
. Words not used in training were also synthesized well.