Rule-based Emotional Voice Conversion Utilizing Three-Layered Model for Dimensional Approach

Japan Advanced Institute of Science and Technology (JAIST)
School of Information Science

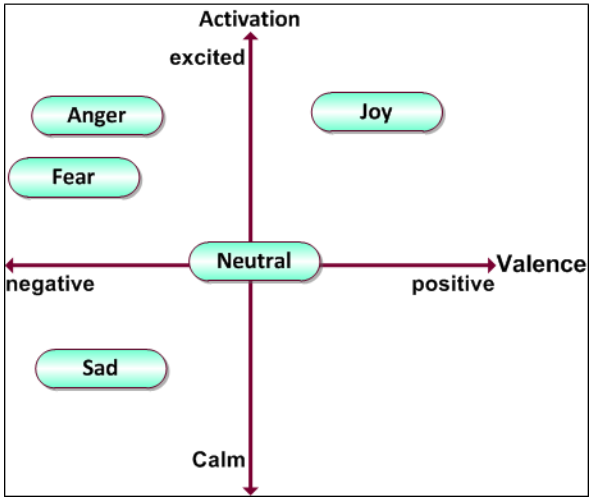
Yawen Xue
Supervisor: Prof. Masato Akagi

Introduction

Category approach is not enough because human can produce emotional speech not only happy but also very happy or a little bit happy

Dimensional approach:

- Representing emotion as a point in a multi-dimensional space
- How negative or positive, how aroused or relaxed can be seen clearly in dimensional space
- Valence: from positive to negative
- Activation: from excited to calm (Grimm et al., Speech Comm)



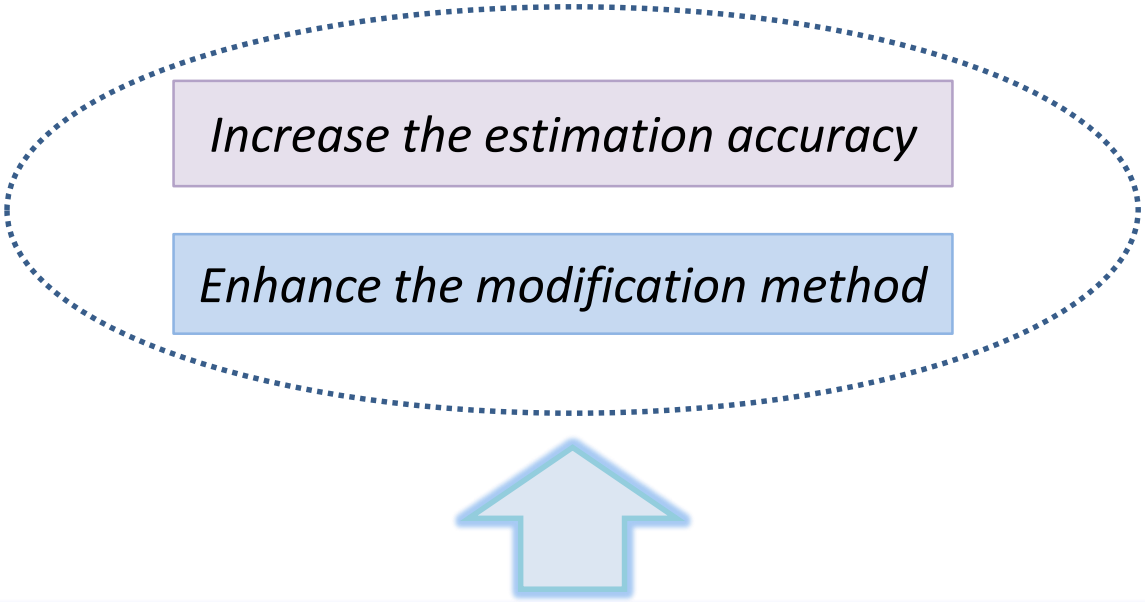
Rule-based synthesis method :

Using the tendency variation of acoustic features which can be acquired with a small database, the synthesized speech can convey all degrees of emotion.

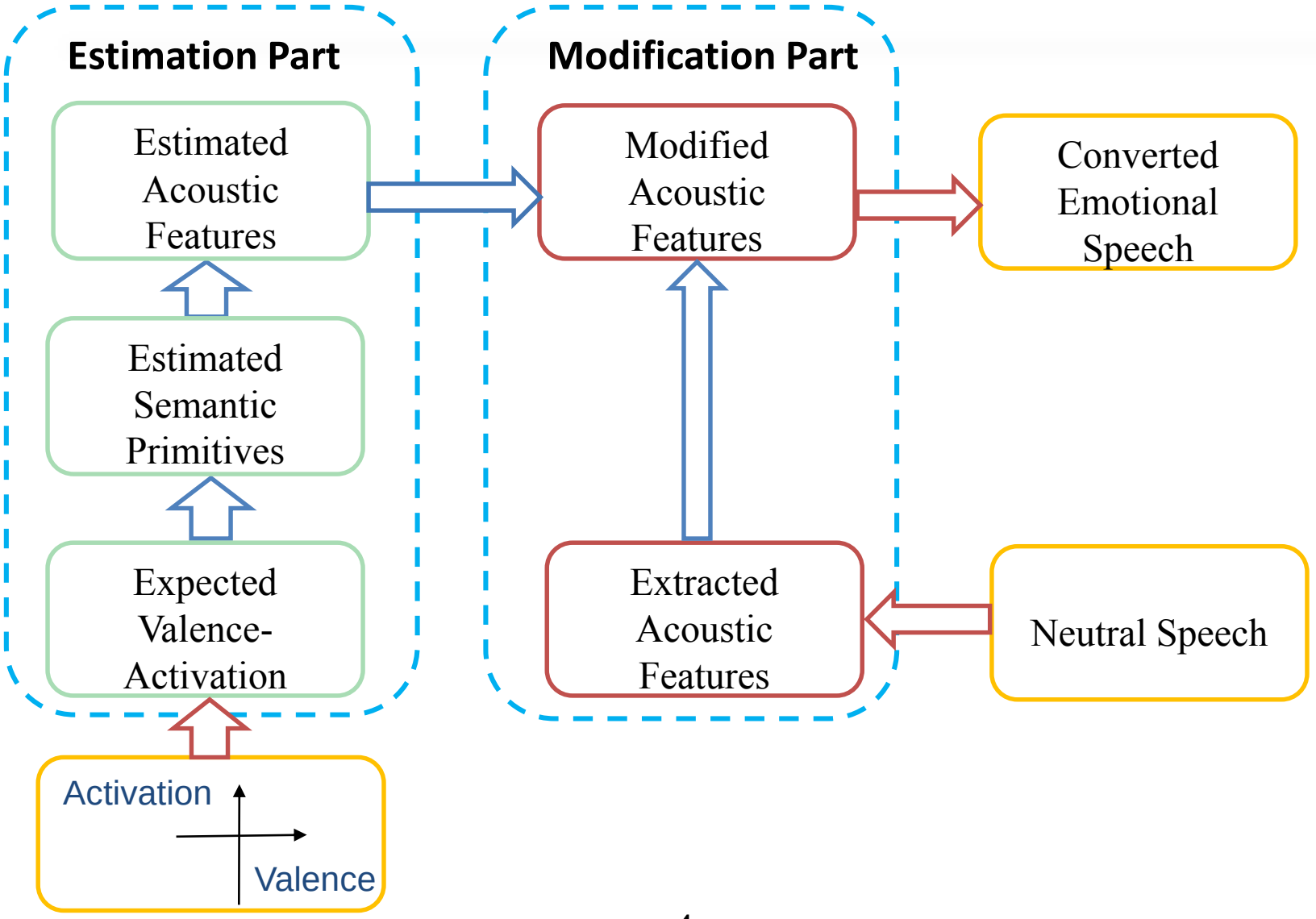
Introduction

Research Purpose

- Propose an emotional speech conversion system based on the three-layered model using dimensional approach

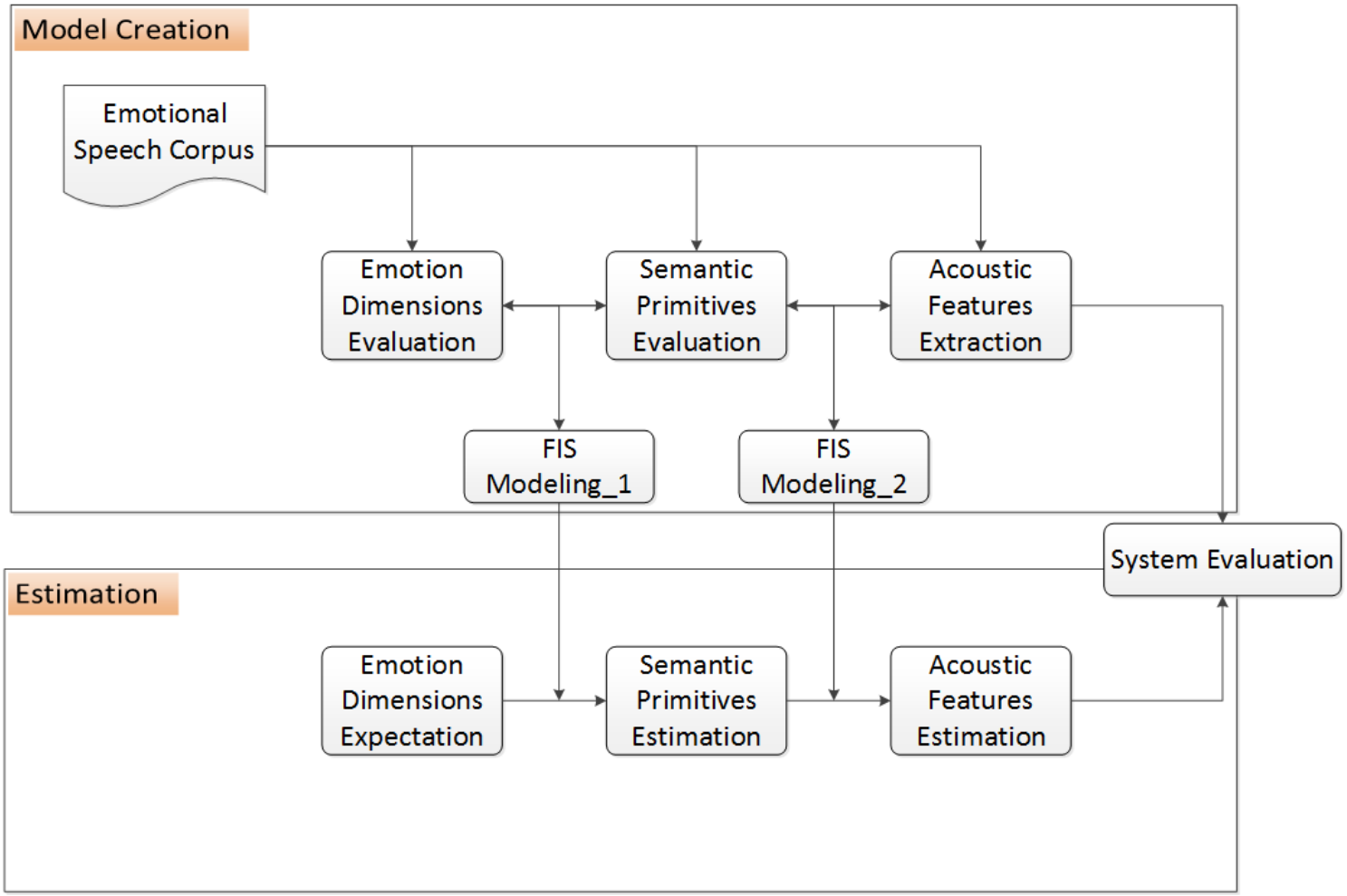


Outline / Emotional Voice Conversion System



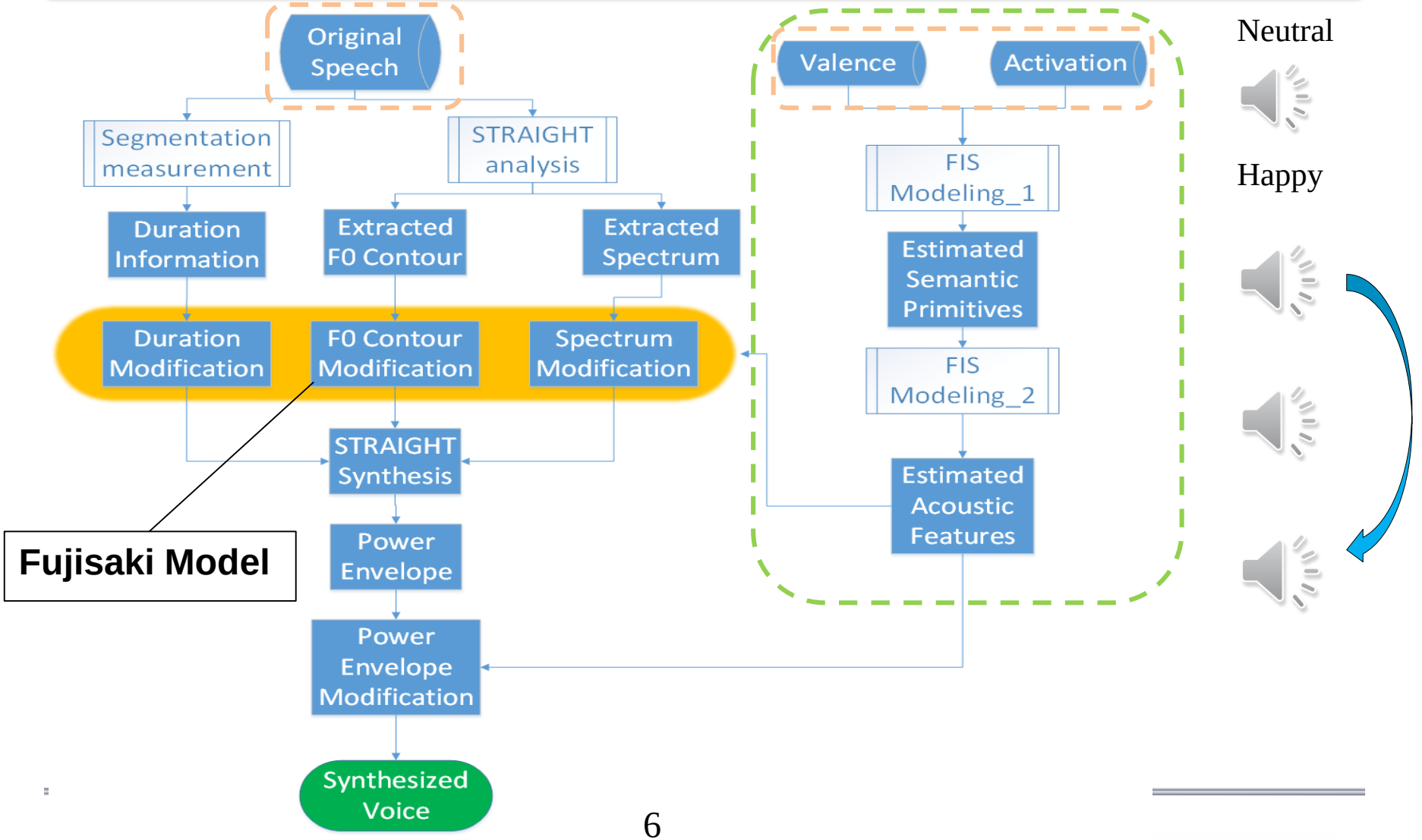
Estimation of Acoustic Features

Two kinds of FIS Modeling are used for estimating the values of semantic primitives and acoustic features



Modification of Acoustic Features

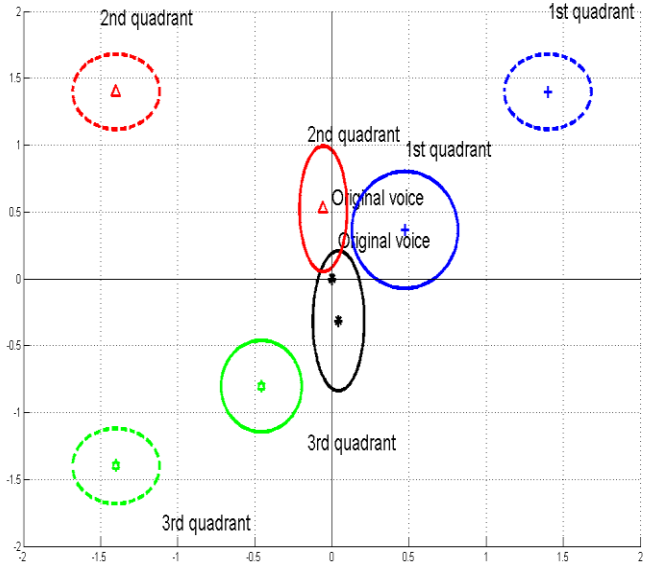
New Point: Fujisaki model is applied for modifying F0 contour (Hamada, Y., Xue Y., & Akagi, M., Wespac 2015)



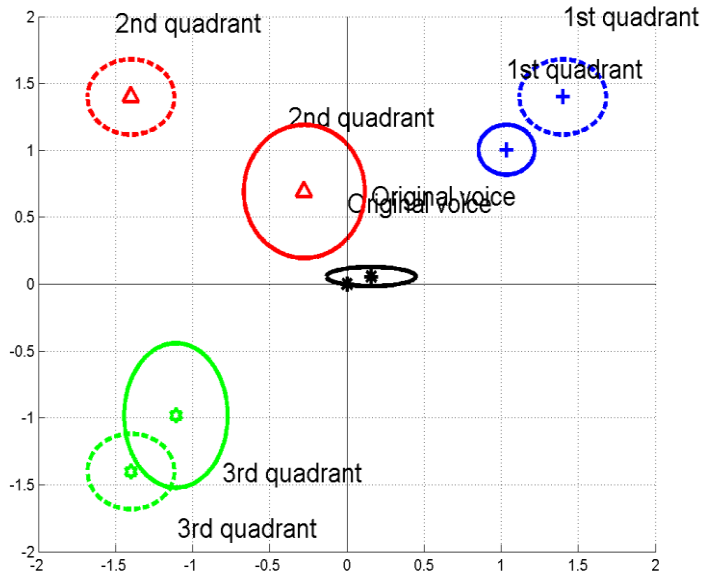
Evaluation Result/ Listening Test

The converted speech from 3-layered model can give the same impression and similar intensity of emotion as intended in the emotion dimension

- **Listening Test is conducted to evaluate the converted speech**
- **Solid: evaluated value for converted voice from listening tests**
- **Dashed: stimulus value for intended emotional voice**



2-layered model



3-layered model

The Closer The Better !

Conclusion and Future Work

Conclusion:

1. The accuracy of estimating the acoustic features is improved using three-layered model.
2. The modification method is enhanced by utilizing Fujisaki model.
3. The converted speech can give the same impression and similar intensity of emotion as intended in the emotion dimension.
4. The converted speech of anger is not ideal comparing with two other emotions.

Future work:

1. As power envelope is much related to anger speech, the method for power envelope modification will be researched in the future

2. Other information from speech such as para-linguistic information will be considered later to build an affective speech synthesis system