Dynamic Subtitle Authoring Method Based on Audio Analysis

Wootaek Lim
Researcher
Realistic Broadcasting Media Research Department

ETRI
INDEX

1. Introduction

2. Proposed Scheme

3. Experiment Result

4. Conclusion
1. INTRODUCTION

• Motivation

• Related Studies
Today, there are 650 million people with disabilities in the world.

More than 66 million people suffering from hearing impairment

They have difficulty in video content understanding due to the loss of audio information.
The main method to make media programs accessible to them is providing subtitles with the video.

The conventional subtitle only provides a basic function to display the actors dialogue.

*People can not understand the context of the scene*

**Examples** of the many type of subtitles used with video content
Related Studies

- The font modifying technic, giving the ambient of scene for hearing impaired

- Express voice volume, speed and pitch by modifying text size, length and height

A scheme to enhance the video accessibility using a “Dynamic Captioning” method

It explores a rich set of technologies including face detection and recognition, speech-script alignment, etc.

2. PROPOSED SCHEME

- Dynamic Subtitle Authoring Method
  - Speech Feature Extraction
    1. Short Time Energy
    2. Pitch
    3. MFCC
  - Speech Feature-Subtitle Matching
  - Speaker Recognition
From the audio signal, speech features are extracted
calculate the voiced frame and align the subtitle optionally
Match the previously extracted features to the subtitle and change it dynamically
Speech Feature Extraction

- Magnitude of each signal frame = Voice volume
- Voice volume impart important information about human emotions
  - *Symbolize the speech energy as a form of emotional information*

- Pitch is the fundamental frequency in a speech, musical note or tone
  - *When human speaks, the fundamental frequency of speech can vary from 100 to 500 Hz*
  - Some studies indicate voice pitch is related to human emotional states
Characteristic of the human ear’s critical bandwidth with frequency

The MFCC makes use of two types of filter based on human ear characteristics: linearly spaced filters and logarithmically spaced filters. *(Linear frequency spacing below 1000 Hz & Logarithmic spacing above 1000 Hz)*
After conducting Mel-frequency filter wrapping, the log mel-spectrum has to be converted back to time domain using DCT.

This result, MFCC is a good representation of the spectral properties of the speech.

MFCC is primarily used as a valid feature in speech/speaker recognition.
**Speech Feature-Subtitle Matching**

- Match the speech energy level and energy variation to the text size and thickness of the letters.

- The fundamental frequency of speech can vary from 100 Hz for male voices to 500 Hz for female voices.

- High-pitched voice to red color & low-pitched voice to violet color.
For speaker recognition, the Gaussian Mixture Model (GMM) was trained using the extracted MFCCs.

GMM is a parametric probability density function represented as a weighted sum of Gaussian component densities.

It is commonly used method for speaker recognition system.

Speaker recognition process is conducted in only script region.
3. EXPERIMENT RESULT

- Experiment result
Test contents are sampled at 16kHz

Frame size for feature extraction is 50ms with 25ms overlap (50% overlap)

Test contents are composed of diverse genres

News, drama, sports and documentary (5~10 min in each)

Number of tested contents = 10 contents

Number of trained speakers (database) = 20 person
Examples of dynamic subtitles based on audio analysis

- Pitch = color
- Energy = size
- Energy variation = bold

Result example #1
Examples of speaker recognition in dynamic subtitles

- Red circle = recognized character
- The accuracy of the speaker recognition in our test is 85%
- Only 10 contents and 20 characters.
4. CONCLUSION

• Conclusion
This paper describes a dynamic subtitle authoring method by analyzing the speech signal.

Extracting the speech features such as STE, ZCR, Pitch and MFCCs.

The major goal is creating the dynamic subtitle for the hearing impaired to understand the media contents better.

It can readily applied to conventional subtitle production system.

The proposed method can be achieved with simple algorithm.
Thank you!