Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds

Kun-Yi Huang, **Chung-Hsien Wu**, Qian-Bei Hong,
Ming-Hsiang Su and Yi-Hsuan Chen

Department of Computer Science and Information Engineering,
National Cheng Kung University, TAIWAN

Outline

- Introduction
- Database
- Proposed Methods
- Experimental Results
- Conclusions

Introduction

- Speech Emotion Recognition (SER) is a hot research topic in the field of Human Computer Interaction. It has a potentially wide applications, such as chatbots, banking, call centers, car board systems, computer games etc.
- In the past, research on speech emotion recognition mainly focused on discriminative emotion features and recognition models.
- Only few existing emotion recognition systems focused on nonverbal part of speech in speech emotion recognition.
  - In real-life communication, nonverbal sounds, such as laughter, cries or emotion interjections, within an utterance play an important role for emotion recognition.
- This work adopted the nonverbal parts to improve the performance of emotion recognition

Goal

- Develop a speech emotion recognition mechanism that considers **verbal** and **nonverbal parts** of speech signals.
- Issues to be considered
  - Emotion database
    - **A spontaneous speech emotion corpus** containing emotional nonverbal sounds in speech
  - Recognition unit
    - **Speech/sound segment** useful to characterize emotion information
  - Temporal Change of Emotion
    - **A sequential model** (seq2seg) for characterizing the temporal change of emotions in a conversation

Literature Review – Emotion Database

Name	Language	A/S	Data	Label
eNTERFACE [E Douglas-Cowie et al.]	English	Acted	Audio, Video	Discr.
EmoDB [F. Burkhardt et al.]	German	Acted	Audio	Discr.
IEMOCAP [C. Busso et al.]	English	Acted& Spont.	Audio, Video, MOCAP	Discr.
RECOLA [F. Ringeval et al.]	French	Spont.	Audio, Video, ECG, EDA	Conti.
CHEAVD [Y. Li et al.]	Chinese	Spont.	Audio, Video	Discr.
NNIME [H. C. Chou et al.]	Chinese	Spont.	Audio, Video, ECG	Discr. & Conti.

- **NNIME, a spontaneous speech emotion corpus**, containing emotional nonverbal sounds in speech, was used for this study.

Literature Review – Recognition Unit

Segment unit	Audio unit	Data	Description
Frame/phoneme/word/utterance	Turn	IEMOCAP, English	Segment based SER using RNN [Tzinis et al., 2018]
Sentence/Second	Turn	IEMOCAP, English	Attentive CNN based SER with different length, features, type of speech [Neumann et al., 2017]
Prosodic action unit	Sentence	English	SVM based SER with discrete intonation patterns [Cao et al., 2014]
Sentence/Word/Syllable	Sentence	IITKGP-SESC, Telugu	SER with local and global prosodic features [Sreenivasa Rao et al., 2012]

- **Discrete prosodic phenomena** can provide complementary information in prediction of emotion. [Cao et al., 2014]

Literature Review –Recognition Model

Method	Input feature	Language	Year
SVM	Prosodic feature	Telugu	[K. S. Rao et al., 2013]
Split vector quantization + naive Bayes	Bag of Audio Words representation	German	[F. B. Pokorny et al., 2015]
Bidirectional LSTM	CNN-extracted vector	French	[G. Trigeorgis et al., 2016]
Attentive CNN	Log-Mels, MFCCs, eGeMAPS	English	[N. T. V. Michael Neumann et al., 2017]
CLDNN	Log-Mels, MFCCs	English	[C.-W. Huang et al.,2017]

- **A sequential model** (seq2seg) is helpful for characterizing the temporal change of emotions in a conversation

Problem –Recognition Unit

- **Problem**
  - Appropriate emotion unit of emotion expression should have various length for recognition. [Tzinis et al., 2018]

- **Proposed method:**
  - We segment the raw audio input utterances with prosodic features as basic emotion unit, which is regarded as a prosodic phrase (PPh).

Problem – Nonverbal Interval Extraction

- **Problem**
  - Non-verbal part of an utterance is helpful for human to recognize emotion.

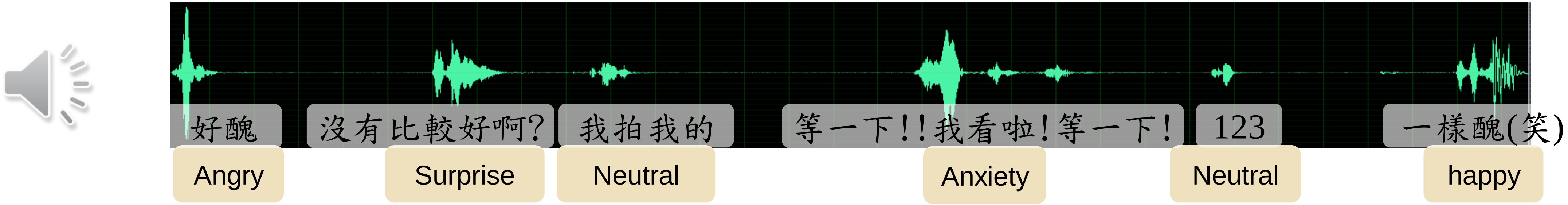
- **Proposed method:**
  - Define **sound types**, such as shout, breath(sobbing), …
  - **Segment** speech utterance into verbal and nonverbal segments.
  - **Extract** sound type features

Problem – Emotion Change in a Conversation

Problem:

- There are **different degree of emotion expression in different time periods** within a speaking turn, so it should be a sequential emotion result to characterize an utterance.

[情境]語者正在試用朋友新買手機的拍照功能



Proposed method:

- We extract **emotion type** and **sound type features** for each segment of input utterance.
- Use **LSTM-based Seq-to-Seq model** to obtain sequential emotion recognition result.

Corpus –NNIME Speech Database

- **NNIME** (NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus)
  - Audio, video, and ECG data
  - Spontaneous emotional speech
  - Recorded by 44 speakers
  - 6 types of **emotion scenario**, 101 sessions, 673.02 mins (11.22 hrs)

Emotion type	Angry	Frustration	Happy	Neutral	Sad	Surprise
Number of sessions	15	19	15	18	18	16

  - Example of scenario setting

Emotion: Angry
Scenario setting: Before going out in the morning, the woman wanted to clean the house while the man was in a hurry. Later, the woman delayed again because she lost some stuff. The man was very angry while the woman was also mad with the man’s temper.

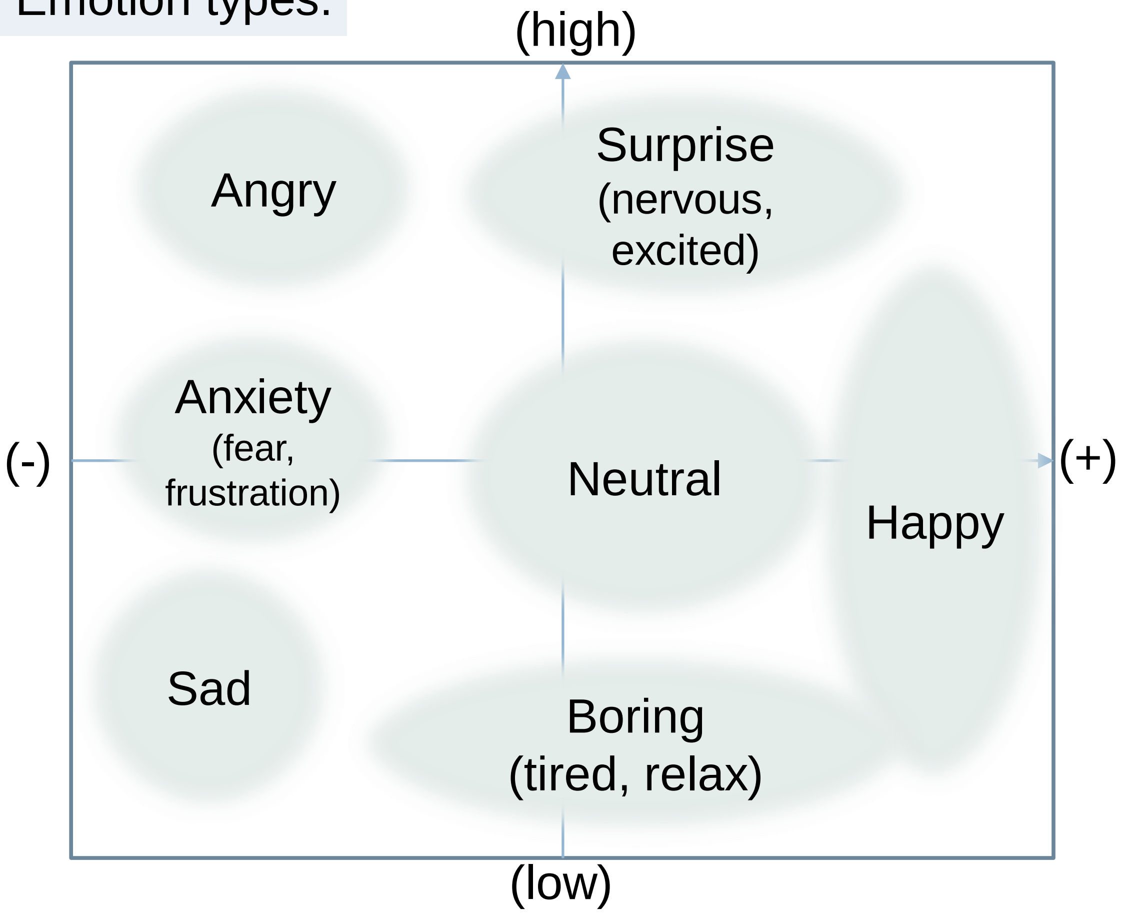
Data Analysis

- **Verbal data**
  - 7 types of emotions
- **Nonverbal data**
  - 3 human sound types+ silence

Sound Type	Description
Shout	shout, scream, howl
Laughter	laugh, giggle
Breathing	sigh, yawn, sob, respire
Silence	silence, noise, audience sound
Verbal	speech

Nonverbal

Emotion types:



Data Statistics

- We segmented all sessions in NNIME into **4766** single speaker dialogue turns.
- Number of segments:14636, duration = 4.3hr (15492.5 secs, $\mu = 3.25,\ \sigma = 5.42$).
  - **All**

Emotion type	Anger	Anxiety	Sadness	Surprise	Neutral	Boring	Happy	Total
Segment number	900	1090	415	1136	5212	537	753	14636

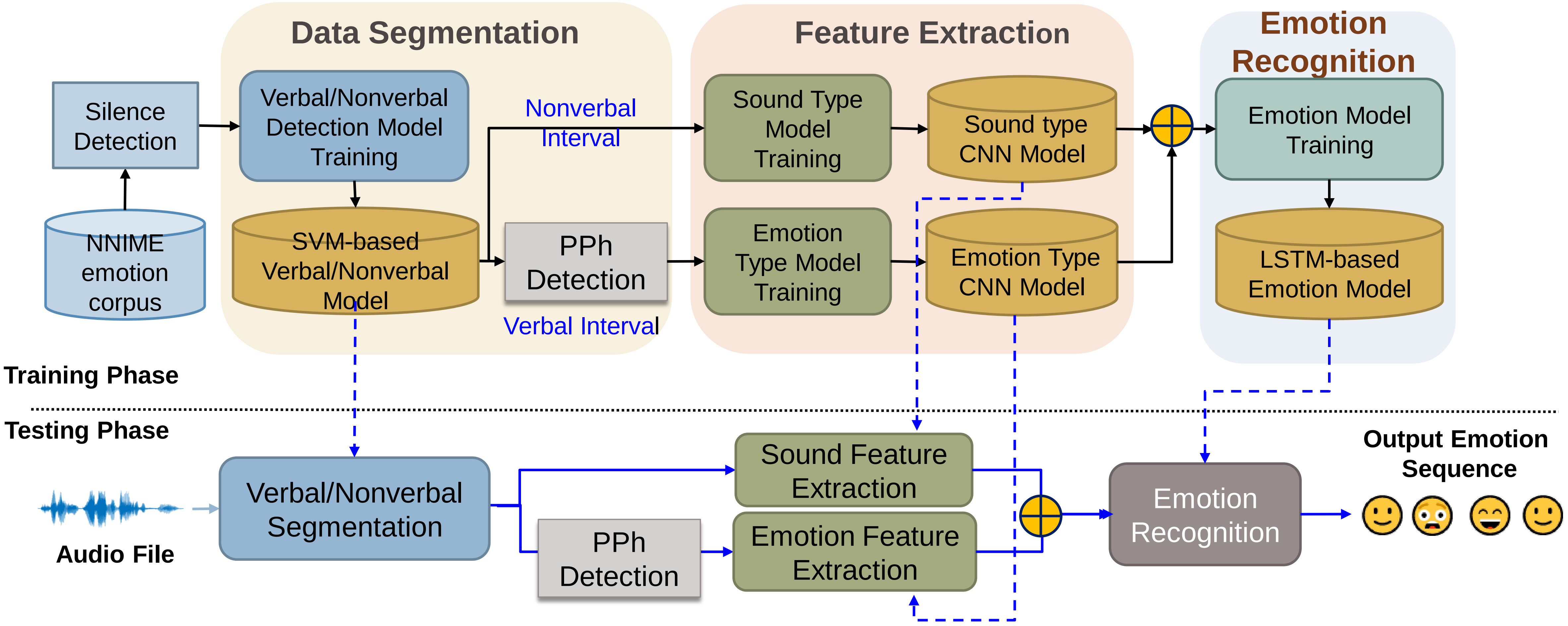
  - **Verbal segments**

Emotion type	Anger	Anxiety	Sadness	Surprise	Neutral	Boring	Happy	Total
Segment number	863	1032	317	1068	5080	491	533	9384

  - **Nonverbal segments**

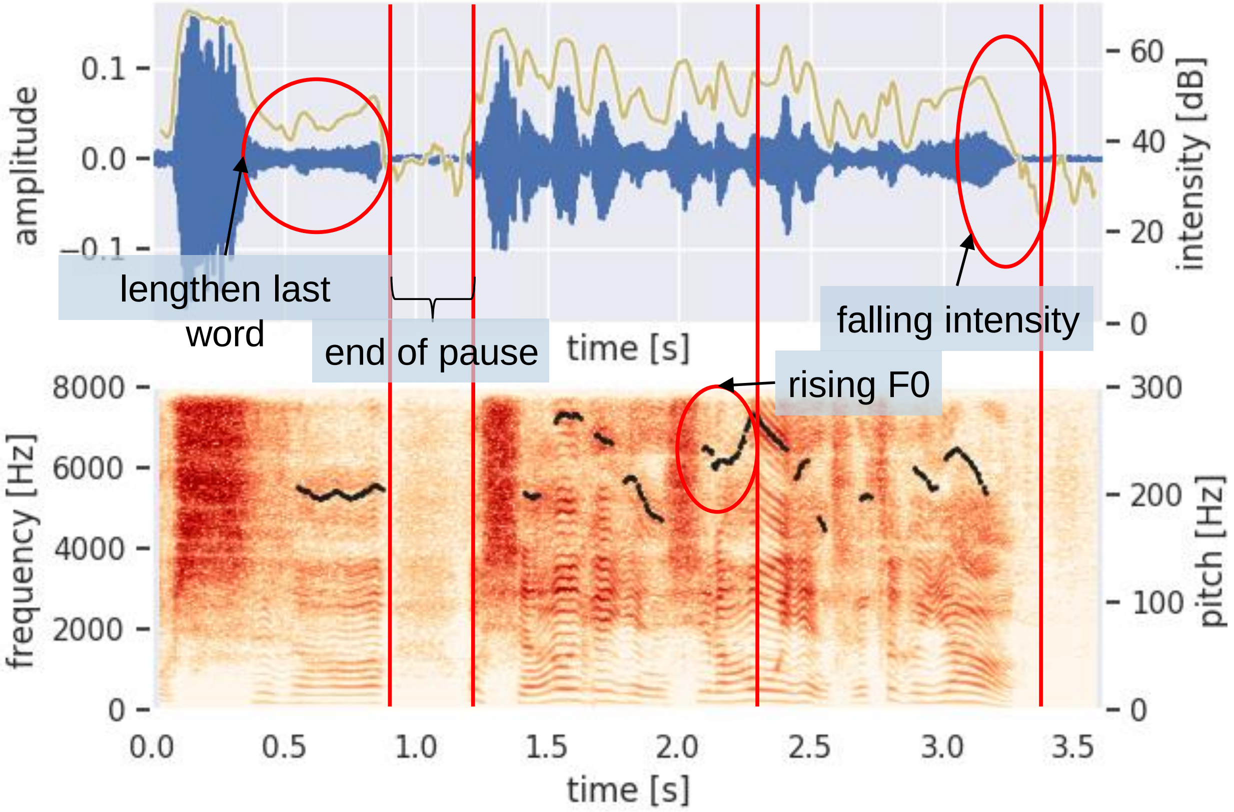
Sound type	Laugh	Breath	Shout	Silence	Total
Segment number	183	409	67	4593	5252

System Framework



Prosodic Phrase Annotation

- Annotate Prosodic Phrase based on the following criteria using *Praat* :
  - Pause (silence for more than 0.3 second)
  - Final rising intonation (Rising F0)
  - Lengthening of last word
  - Sharp fall in intensity (Falling intensity)
  - Modified wrong annotation of silence interval



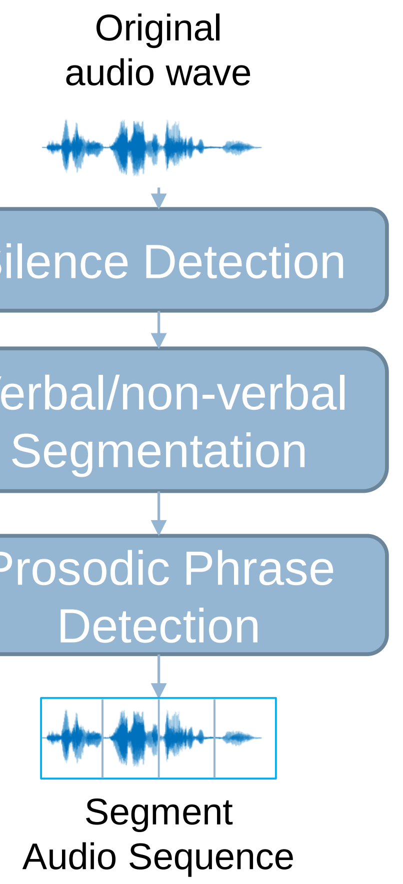
Audio Data Segmentation

- **Silence interval detection**: produced by *Praat*
- **Verbal/ Non-verbal Segmentation**:
  1. Extract frame-based 384-dim audio feature by *openSMILE* [F. Eyben et al.]
  2. Calculate probability sequence of verbal/non-verbal frames by SVM
  3. Smoothing the probability sequence and compute boundary score

$$\delta(P) = \left| \sum_{i=1}^{3} (4-i)^2 * P[b-i] - \sum_{i=1}^{3} (4-i)^2 * P[b+i] \right|$$

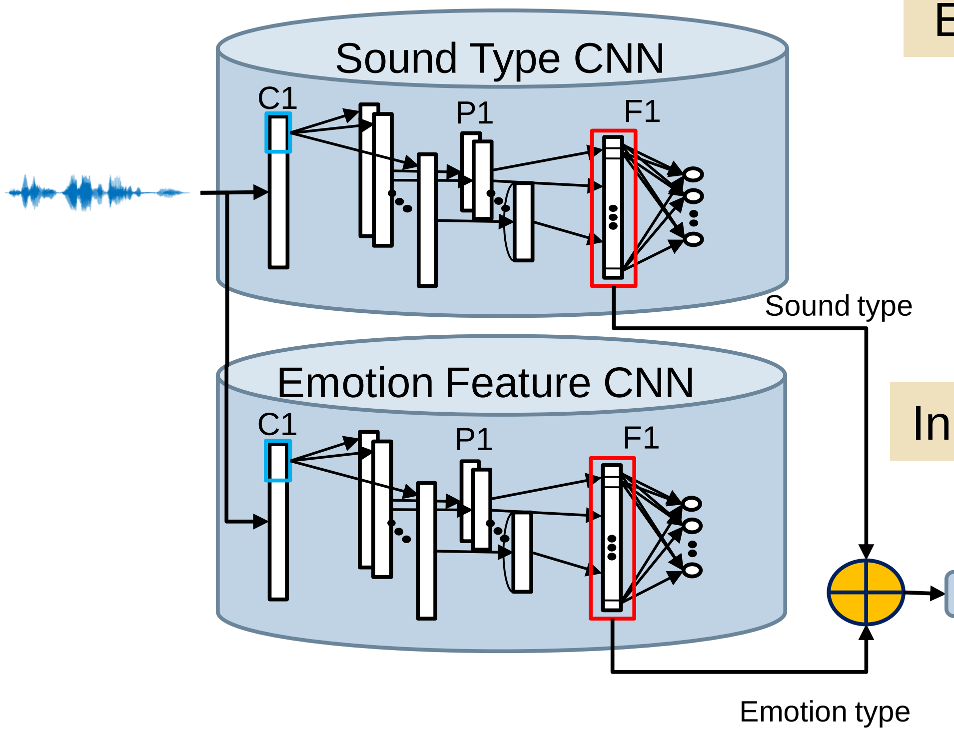
  4. If boundary score > threshold, set it as a boundary.
- **Prosodic Phrase Detection**: PPh detected by *PPh Autotagger*

[Domínguez et al., 2016a]



Feature Vector for each Segment

- Using raw waveforms as input of CNN. [Bertero et al., 2017
- 4 sound types and 7 emotion types
- The **last hidden layer output** is used as feature vector for recognition.



Emotion feature/Sound type Vector

$$\begin{cases} Ei: the\ hidden\ vector\ of\ emotion\ feature\ extraction\ CNN \\ Si: the\ hidden\ vector\ of\ sound\ type\ classification\ CNN \end{cases}$$

$$i = \begin{cases} the\ i_{th} \begin{cases} pure\ verbal\ segment \\ pure\ nonverbal\ segment' \end{cases} & in\ training\ phase \\ the\ i_{th}\ segment\ of\ input\ audio, & in\ testing\ phase \end{cases}$$

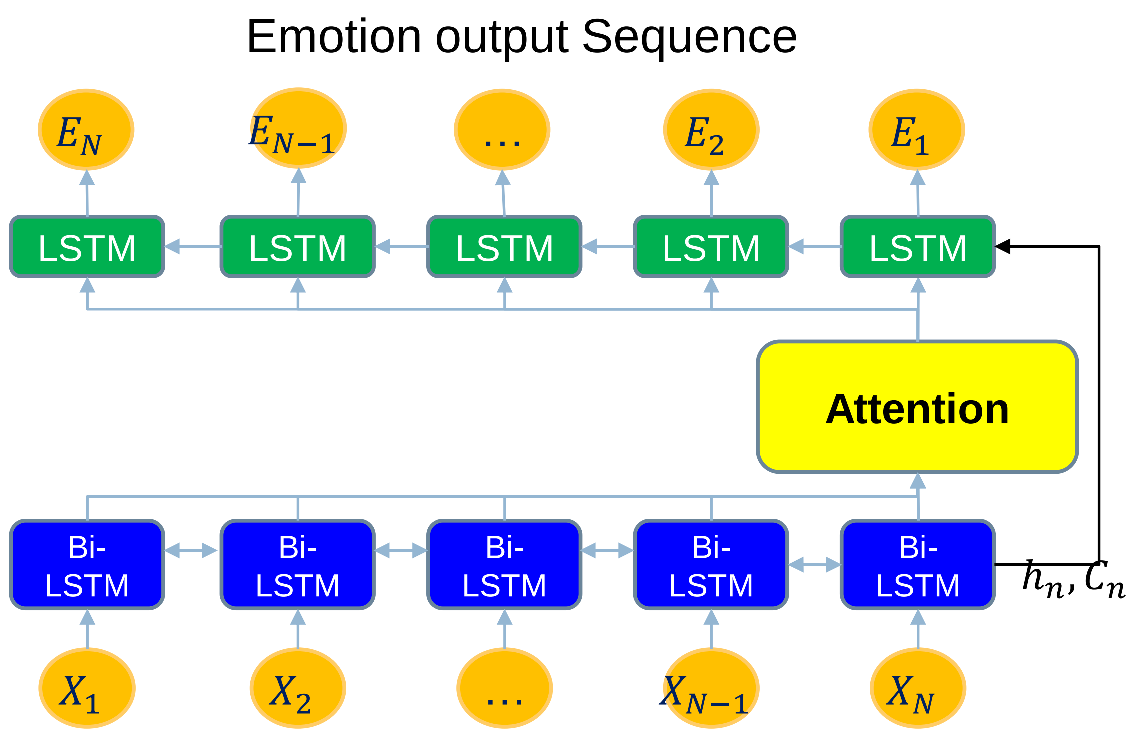
Input Vector for Seq2seq Emotion Recognition

$$X_i = S_i \oplus E_i$$

Feature vector of audio segment → Seq2seq Emotion Recognition Model

Attentive Bi-LSTM based Seq-to-Seq Model

- The sound features for nonverbal segment and emotion features for each segment were adopted as feature vector X_i to feed to the **LSTM based Seq-to-Seq emotion recognition model with attention.**



$$X_i = S_i \oplus E_i \quad i = 1, \dots, N$$

N = number of segments in the utterance

Experimental Results - Evaluation on verbal/nonverbal segmentation

- 300 dialog turns from each pre-specified emotion and duration range were manually labeled for evaluation
  - Features with dimensionalities of 32 and 384 were selected with window sizes of 100ms and 200ms and a shift size of 50ms.
  - A boundary is labeled correctly if the detected label is within **100ms** of the manually labeled time.
  - The precision, recall, F1 score was used for evaluation
    - F is feature dimension
    - W is the window size
    - S is the shift size
    - FM (100ms) is full match
    - PM (200ms) is partial match

		F = 32 W = 100 S = 50		F = 32 W = 200 S = 50		F = 384 W = 100 S = 50		F = 384 W = 200 S = 50	
		FM	**PM**	**FM**	**PM**	**FM**	**PM**	**FM**	**PM**
0.6	**Pre**	0.24	0.46	0.23	0.47	0.34	0.62	0.31	0.57
	Rec	0.31	0.60	0.29	0.58	0.37	0.65	0.36	0.66
	F1	0.27	0.52	0.25	0.51	0.35	0.63	0.33	0.61
0.8	**Pre**	0.24	0.46	0.23	0.47	**0.37**	**0.66**	0.32	0.53
	Rec	0.31	0.60	0.28	0.57	**0.37**	**0.64**	0.36	0.64
	F1	0.27	0.52	0.25	0.51	**0.37**	**0.64**	0.34	0.61
1	**Pre**	0.25	0.48	0.23	0.49	0.38	0.67	0.33	0.59
	Rec	0.30	0.59	0.27	0.56	0.35	0.60	0.35	0.62
	F1	0.27	0.53	0.25	0.51	0.36	0.63	0.34	0.60
1.2	**Pre**	0.26	0.50	0.23	0.50	0.41	0.69	0.35	0.61
	Rec	0.30	0.58	0.26	0.55	0.32	0.54	0.34	0.58
	F1	0.28	0.54	0.24	0.52	0.36	0.61	0.34	0.59

Experimental Results - Evaluation on Feature Extraction

- This work selected a number of filters and different sizes in the adaptive pooling layer based on the accuracy of emotion classification
- The results of comparison between the methods using raw speech signal and extracted acoustic feature sets were obtained
- Performance of emotion type classification

Input	Best parameters	Accuracy
Speech signal	Filter number = 100, Kernel size = 512, step = 256, pooling = 2	30.10%
32-dim LLDs	Filter number = 150, Kernel size = 2, step = 1, pooling = 2	26.10%
32-dim LLDs with 12 functionals	Filter number = 100, Kernel size = 2, step = 1, pooling = 10	21.20%

Experimental Results - Evaluation of Feature Extraction

- Performance of sound type classification

Input	Best parameters	Accuracy
Speech signal	Filter number = 100, Kernel size = 512, step = 256, pooling = 2	54.90%
32-dim LLDs	Filter number = 100, Kernel size = 2, step = 1, pooling = 2	53.63%
32-dim LLDs with 12 functionals	Filter number = 250, Kernel size = 2, step = 1, pooling = 10	47.95%

- The last hidden layer outputs of the CNN emotion/sound models were concatenated and fed to the LSTM-based sequence-to-sequence model for emotion recognition

Experimental Results - Evaluation of Emotion Recognition

- The hidden layer sizes of the LSTM were selected from 32, 64, 128, 256, and 512 to achieve the highest accuracy of emotion recognition
  - The proposed method achieved 52.00% when the hidden size of the LSTM was set to 128
- This work compared the performance of the proposed method with traditional emotion recognition models with frame-based acoustic features or raw speech signal as input

	Input	Best parameters	Accuracy
Proposed method	CNN-based feature extraction	Hidden size = 128	52.00%
LSTM	32-dim LLDs	Hidden size = 256	44.30%
CNN	Speech signal	Pooling = 2, filter number = 100	30.10%

Conclusion and Discussion

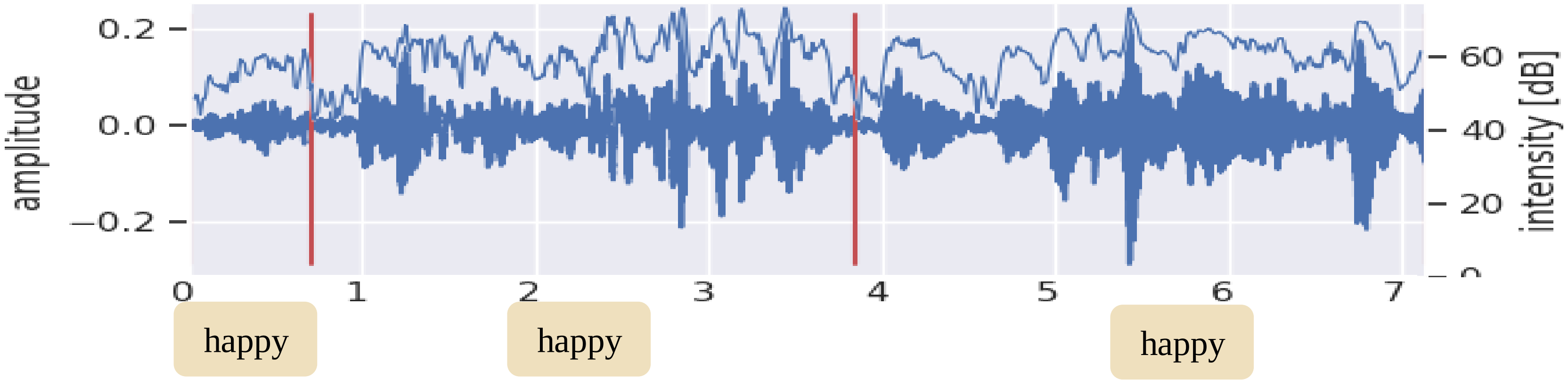
Conclusion

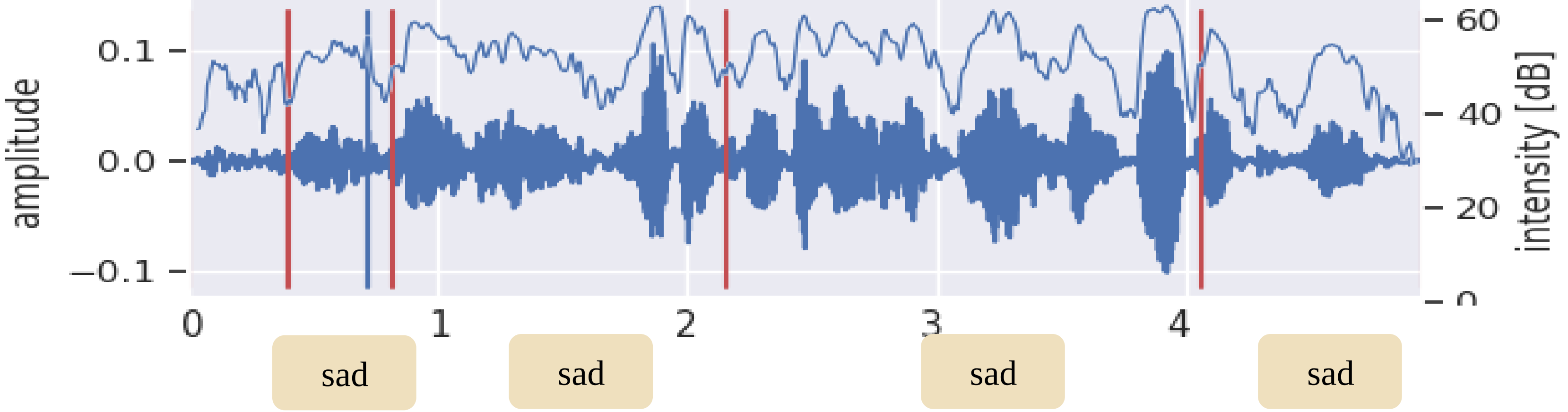
- Speech emotion recognition considering nonverbal interval and types of sound achieved a better performance.
- Sequence-to-sequence model can characterize emotional change in a dialogue turn.

Discussion

- Emotion expression in spontaneous speech is very diverse and difficult to be labeled with one specific emotion.
- The other difficulty of spontaneous speech emotion recognition is the background noise. Preprocessing of audio data is an important issue.
- There are still many sound types in our daily conversation. The types of emotional sound event should be better defined.

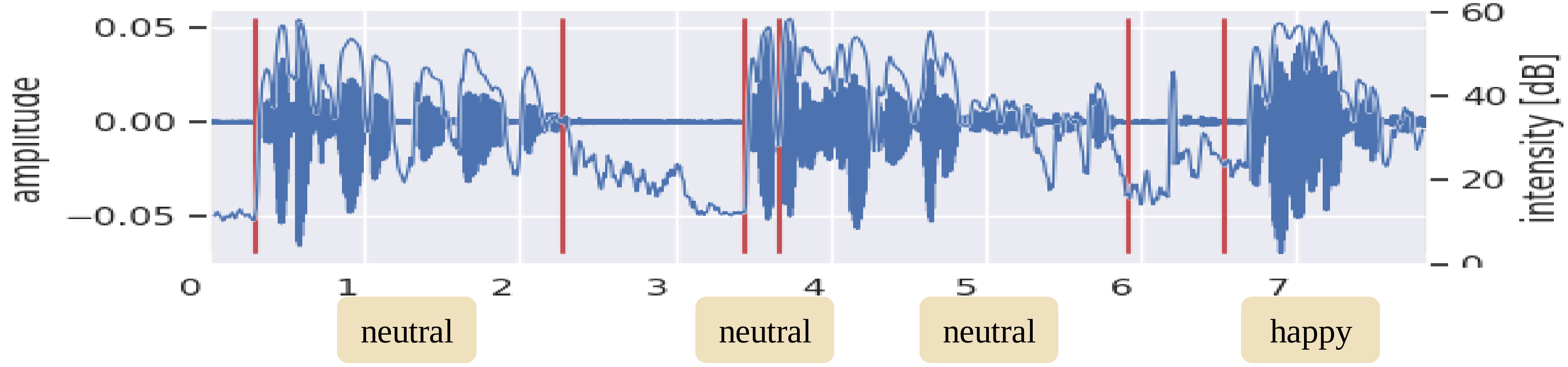
Result Demo – Inside

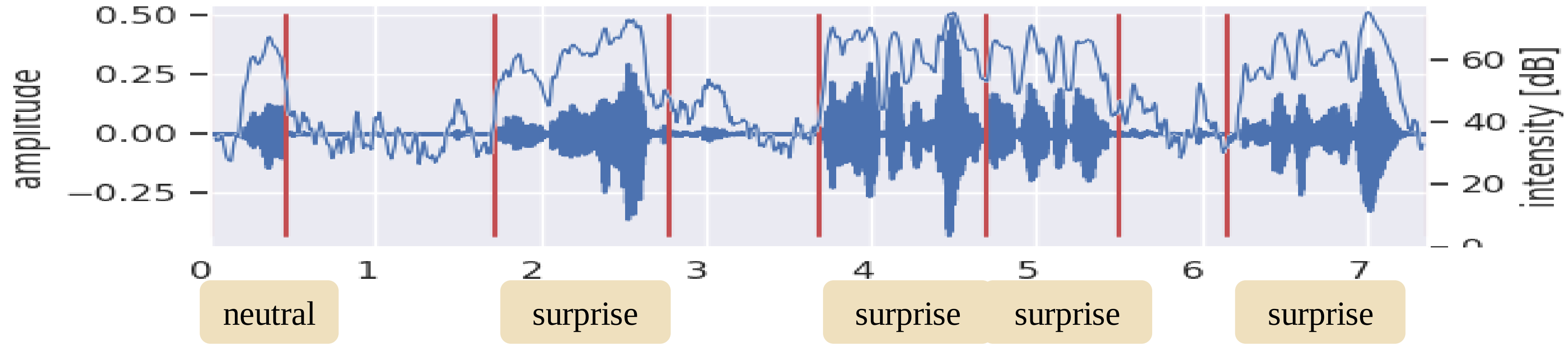




Result Demo – Outside

- These audios are from NNIME sessions which are used for training.





Questions?