Automatic Naturalness Recognition from Acted Speech Using Neural Networks

Bagus Tris Atmaja and Akira Sasou, AIST Japan
Masato Akagi, JAIST Japan

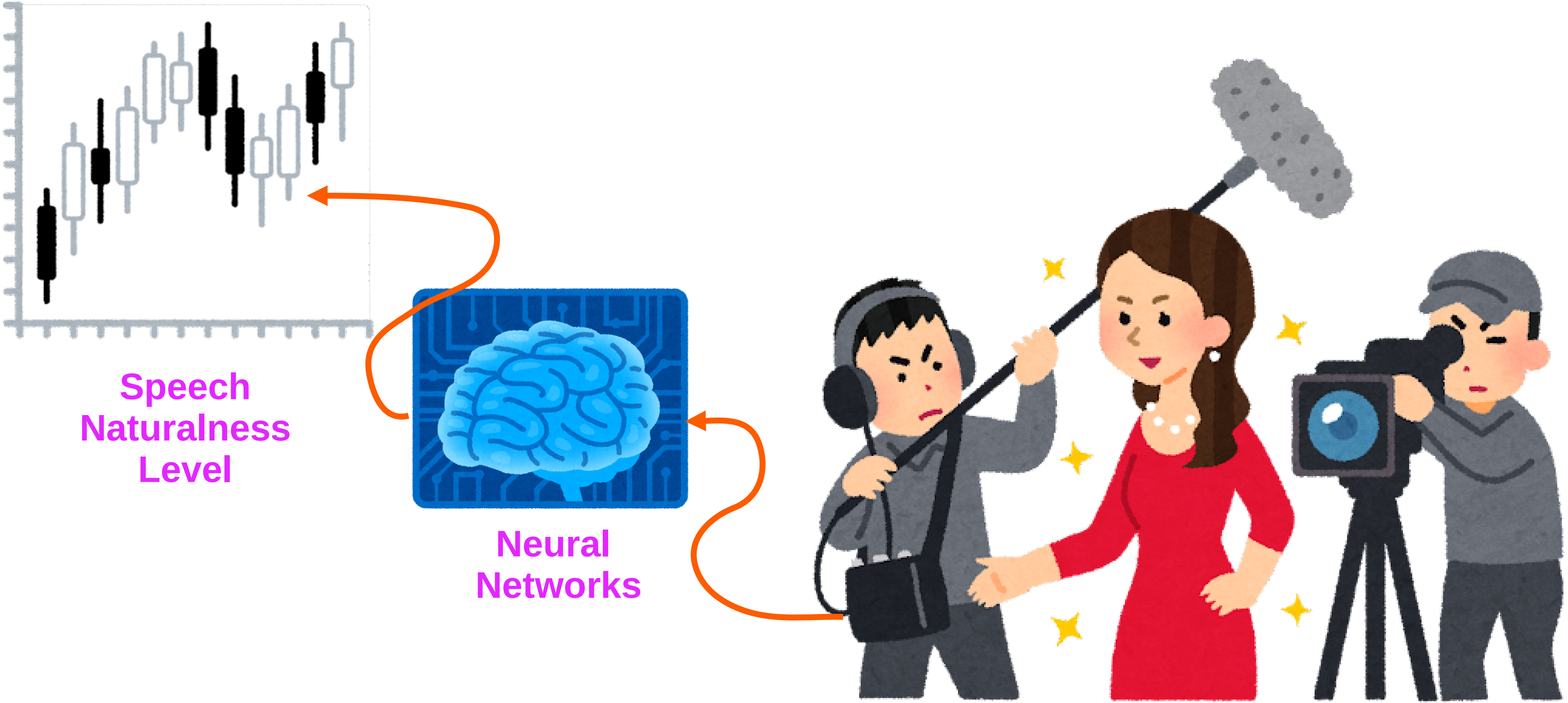
JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
1990

産総研

Motivation

- Speech naturalness recognition is already developed, but is used to measure naturalness in either speech synthesis or in psychophysics (e.g., naturalness for stutterers).
- A new application of speech naturalness recognition can be aimed at measuring naturalness of acted dialogue.
- This study proposes an evaluation of naturalness of speech from acted dialogue via neural network mechanism.

Potential application



Speech Naturalness Level

Neural Networks

Movie making

Problem Statement

1) Given speech utterances (provided in .wav files) from acted dialogues with naturalness recognition labels measured at 5-point scales, is it possible to recognize these labels automatically using neural networks?

2) How to perform automatic naturalness recognition and evaluate the methods?

Dataset: MSP-IMPROV

Characteristics	Acted dialogue, propose naturalness
# Utterances	8438
Scenarios	Target-improvised, Other-improvised, Target-read, and Natural interaction
# Speakers	12
# Sessions	6 (Training: 1-5, Test: 6)
Naturalness labels	1 to 5

Acoustic Features

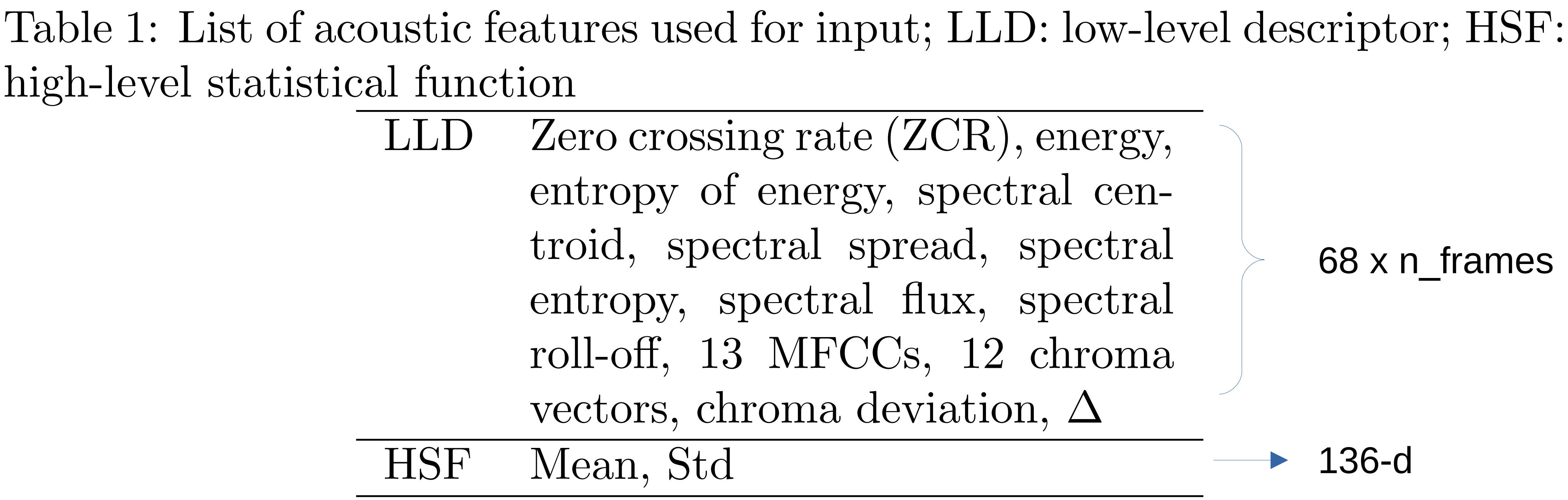


Table 1: List of acoustic features used for input; LLD: low-level descriptor; HSF: high-level statistical function

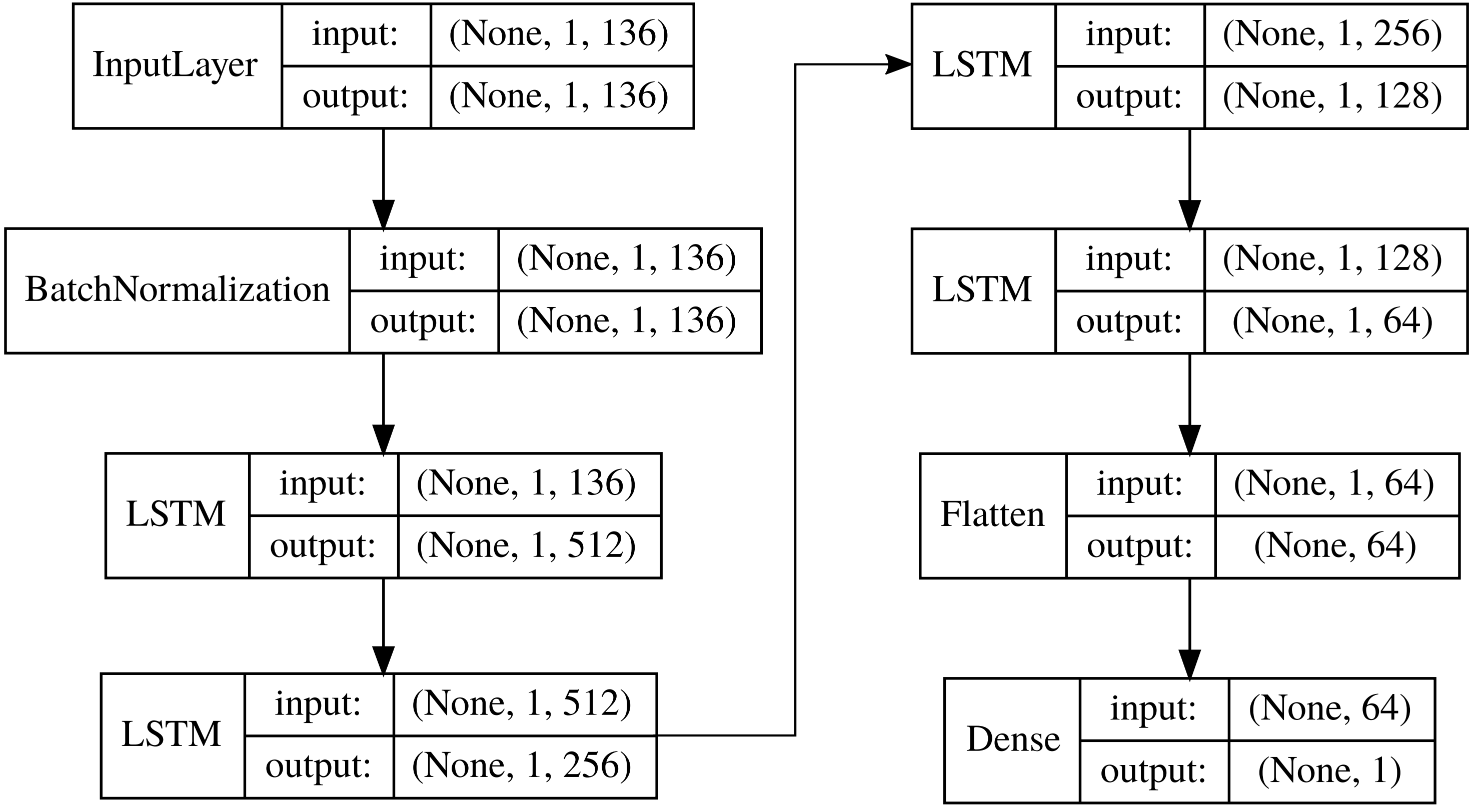
LLD	Zero crossing rate (ZCR), energy, entropy of energy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, 13 MFCCs, 12 chroma vectors, chroma deviation, Δ
HSF	Mean, Std

68 x n_frames

136-d

Classifiers

MLP & LSTM (varying both number of layers and nodes)



Layer		Shape
InputLayer	input:	(None, 1, 136)
	output:	(None, 1, 136)
BatchNormalization	input:	(None, 1, 136)
	output:	(None, 1, 136)
LSTM	input:	(None, 1, 136)
	output:	(None, 1, 512)
LSTM	input:	(None, 1, 512)
	output:	(None, 1, 256)
LSTM	input:	(None, 1, 256)
	output:	(None, 1, 128)
LSTM	input:	(None, 1, 128)
	output:	(None, 1, 64)
Flatten	input:	(None, 1, 64)
	output:	(None, 64)
Dense	input:	(None, 64)
	output:	(None, 1)

Evaluation Metric

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{1}$$

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \tag{2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{n}((x_i - \mu_x)^2 + (y_i - \mu_y)^2)}. \tag{3}$$

Result: MLP with LLD

# layers	# units	CCC	PCC	RMSE
1	16	- 0.005	-0.008	0.334
2	32, 16	- 0.003	-0.025	0.313
3	64, 32, 16	0.106	0.197	0.307
4	128, 64, 32, 16	0.147	0.247	0.305
5	256, 128, 64, 32, 16	0.124	0.184	0.314
6	512, 256, 128, 64, 32, 16	0.115	0.237	0.304
5	512, 256, 128, 64, 32	0.112	0.268	0.301
4	512, 256, 128, 64	0.124	0.200	0.309
3	512, 256, 128	0.139	0.234	0.308
2	**512, 256**	**0.160**	**0.250**	**0.304**
1	512	0.126	0.200	0.310

Result: LSTM with LLD

# layers	# units	CCC	PCC	RMSE
1	16	0.127	0.149	0.438
2	32, 16	0.138	0.169	0.476
3	64, 32, 16	0.211	0.224	0.374
4	128, 64, 32, 16	0.225	0.241	0.356
5	256, 128, 64, 32, 16	0.255	0.260	0.357
6	512, 256, 128, 64, 32, 16	0.115	0.237	0.304
5	512, 256, 128, 64, 32	0.230	0.260	0.367
4	512, 256, 128, 64	0.242	0.247	0.360
3	**512, 256, 128**	**0.269**	**0.274**	**0.357**
2	512, 256	0.143	0.134	0.431
1	512	0.131	0.161	0.343

Result: MLP with HSF

# layers	# units	CCC	PCC	RMSE
1	16	0.215	0.302	0.299
2	32, 16	0.220	0.294	0.302
3	**64, 32, 16**	**0.228**	**0.311**	**0.299**
4	128, 64, 32, 16	0.210	0.286	0.302
5	256, 128, 64, 32, 16	0.203	0.287	0.302
6	512, 256, 128, 64, 32, 16	0.219	0.294	0.302
5	512, 256, 128, 64, 32	0.214	0.293	0.301
4	512, 256, 128, 64	0.213	0.305	0.298
3	512, 256, 128	0.196	0.279	0.302
2	512, 256	0.199	0.284	0.302
1	512	0.197	0.288	0.300

Result: LSTM with HSF

# layers	# units	CCC	PCC	RMSE
1	16	0.258	0.273	0.363
2	32, 16	0.245	0.259	0.363
3	64, 32,16	0.268	0.290	0.347
4	128, 64,32,16	0.245	0.272	0.346
5	256, 128, 64, 32, 16	0.280	0.299	0.360
6	512, 256, 128, 64, 32, 16	0.300	0.314	0.357
5	512, 256, 128, 64, 32	0.284	0.313	0.330
4	**512, 256, 128, 64**	**0.302**	**0.327**	**0.339**
3	512, 256, 128	0.267	0.286	0.355
2	512, 256	0.273	0.299	0.345
1	512	0.274	0.280	0.353

Conclusions

1) The naturalness of acted dialogue can be recognized by such a neural network mechanism; we demonstrate the ability of simple MLP and LSTM networks to predict speech naturalness over different layers and nodes

2) The evaluation of speech naturalness recognition shows moderate performance in terms of concordance coefficient correlation (CCC)