Effect of different splitting criteria on the performance of speech emotion recognition

Bagus Tris Atmaja and Akira Sasou
National Institute of Advanced Industrial Science & Tech., Japan

Motivation

- Traditional speech emotion recognition (SER) evaluations have been performed merely on a speaker-independent condition.
- In a literature review [1], it is suggested that acoustic information depends on linguistic information.
- Changing SER evaluation by splitting different linguistic information may lead to different results
- This paper highlights the importance of splitting training and test data for SER by script, known as sentence-open or text-independent criteria.

[1] H. Fujisaki, “Prosody, Information, and Modeling with Emphasis on Tonal Features of Speech,” in Workshop on Spoken Language Processing, 2003.

Related Work [2]

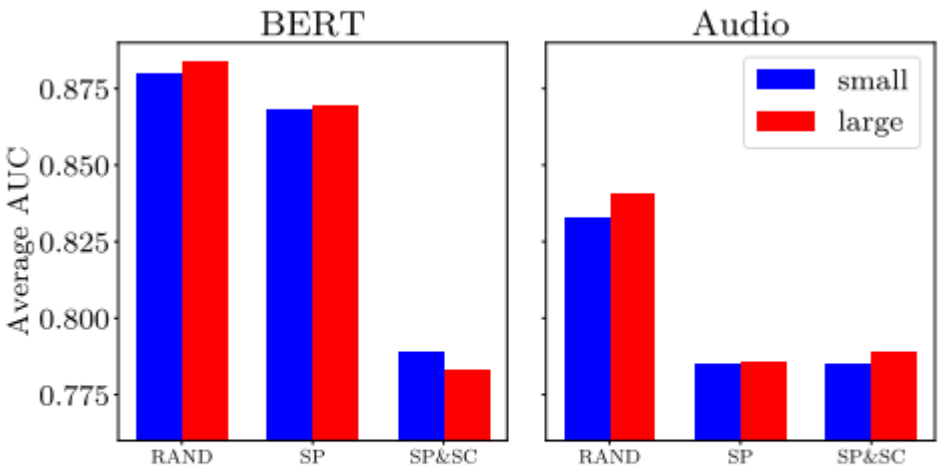


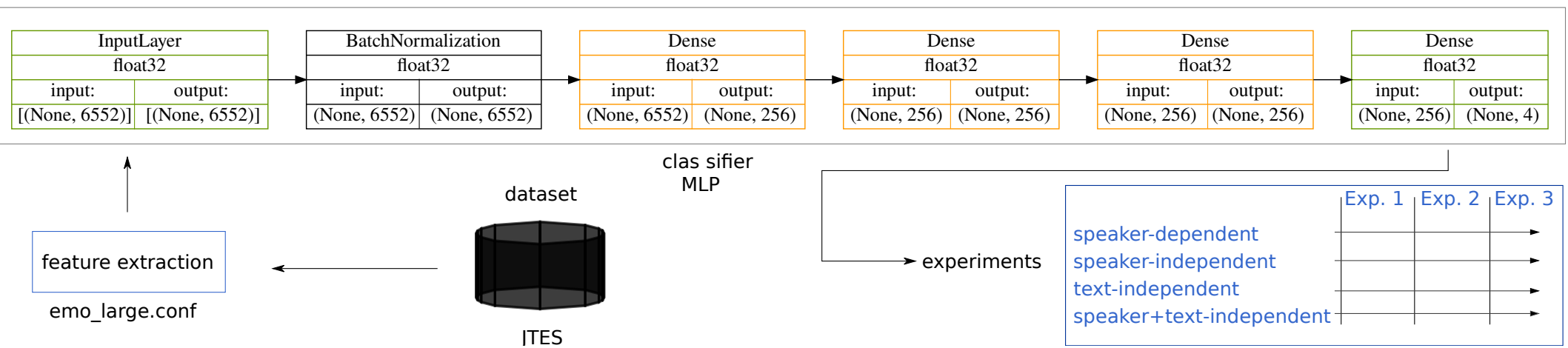
Fig. 2: Effect of different criteria for defining the folds in IEMOCAP on audio- and text-based systems for two different model sizes (small and large). RAND: random folds, SP: by-speaker folds, SP&SC: by-speaker and by-script folds.

While the effect of by-script folds is found on acoustic-linguistic emotion recognition, no study has been conducted to investigate the effect of similar phenomena on acoustic-only emotion recognition (SER)

[2] L. Pepino, P. Riera, L. Ferrer, and A. Gravano, “Fusion Approaches for Emotion Recognition from Speech Using Acoustic and Text-Based Features,” in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6484–6488.

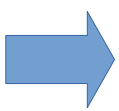
Methods

1) Dataset: JTES (Japanese Twitter-based Emotional Speech) Corpus
2) Acoustic feature: low-level descriptor (LLD) in “emo_large” configuration of OpenSmile toolkit.
3) Classifier: Multilayer Perceptron (MLP)



Splitting Criteria

JTES Corpus



	Training	Test
SD	19600 samples	400 samples
SI	98 speakers	2 speakers
TI	49 sentences	1 sentence
STI	90 speakers x 40 sentences	2 speakers x 1 sentence

SD: speaker-dependent
TI: text-independent

SI: speaker-independent
STI: speaker+text-independent

Experiments

- **Experiment I: Average of 30 trials**
- **Experiment II: Cross-validation**
- **Experiment III: Same-number of test data**

Criteria	Exp. #1 & Exp. #2		Exp. #3	
	Training	Test	Training	Test
SD, SI, TI	19600	400	14400	400
STI	14400	400	14400	400

Result

Results in weighted accuracy (WA) ± standard error (SE) on different experiment conditions

Criteria	WA (%) ± SE (%)		
	Exp. #1	Exp. #2	Exp. #3
SD	91.14 ± 0.07	92.30 ± 0.40	89.40 ± 0.48
SI	87.88 ± 0.09	88.85 ± 0.49	86.64 ± 0.63
TI	64.36 ± 0.08	65.04 ± 0.90	62.35 ± 0.93
STI	69.56 ± 0.09	70.65 ± 0.44	70.65 ± 0.44

Conclusion

- SER with different linguistic information for training and test (i.e., text-independent) is more difficult than other criteria.
- Evaluating SER with text-independent is a challenging task for future research:
  - Enlarge the dataset for more linguistic information coverage
  - Propose special technique to tackle text-dependency