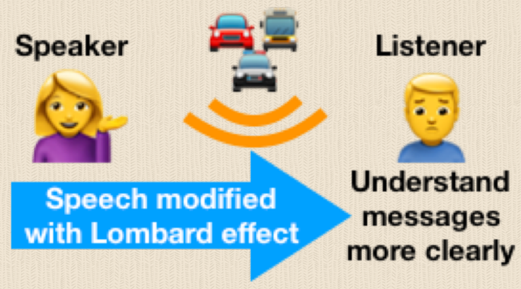
Does the Lombard Effect Improve Emotional Communication in Noise?
– Analysis of Emotional Speech Acted in Noise –

Yi Zhao[1], Atsushi Ando[2], Shinji Takaki[1], Junichi Yamagishi[1], Satoshi Kobashikawa [2]
[1]National Institute of Informatics, Japan [2]NTT Media Intelligence Laboratories, Japan

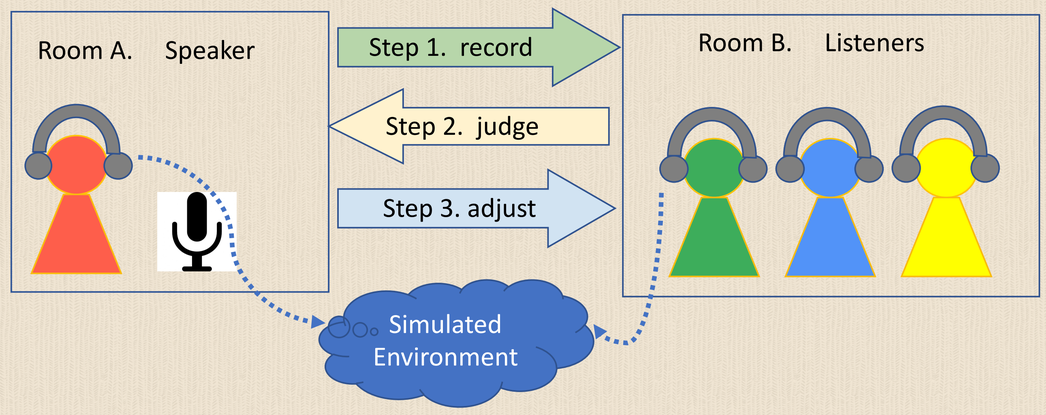
Introduction

- Speakers usually adjust their way of talking in noisy environments involuntarily for effective communication. This adaptation is known as the Lombard effect.



- We investigate how the Lombard effect affects emotional speech:
  1). Can speakers express their emotions correctly even under adverse conditions?
  2). Can listeners recognize the emotion contained in speech signals even under noise?
  3). How does emotional speech uttered in noise differ from emotional speech uttered in quiet conditions in terms of acoustic characteristics?

Recording Procedure



- 12 speakers (6 highly-trained and 6 junior-level), 36 listeners.
- Four acted emotions: happy, sad, angry, neutral.
- At least 80 successfully produced and correctly pronounced utterances, 40 in a quiet environment and 40 in noise.
- Noise: a mixture of speech-shaped noise called ICRA noise and in-car noise (SNR≈-8.7dB).

Analysis based on Confusion Matrix & Frobenius Distance

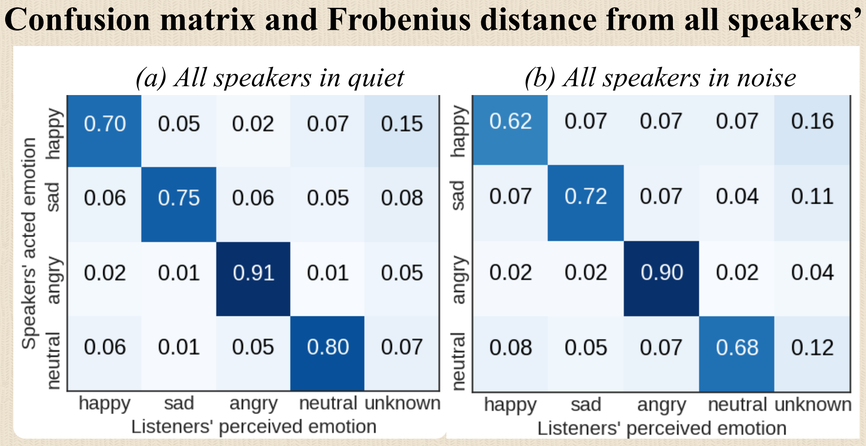
- Frobenius distance:

$$F = \mathbf{Tr}[(\mathscr{C} - \mathscr{I})^{\top}(\mathscr{C} - \mathscr{I})]$$

F: Frobenius distance; $\mathscr{C}$: the confusion matrix except for the 'unknown' column. $\mathscr{I}$: 4×4 identity matrix.

- The smaller the confusability of emotional speech, the smaller the Frobenius distance.

A. Confusion matrix and Frobenius distance from all speakers' perspective



Frobenius distance: 0.172 Frobenius distance: 0.379

- **Speeches uttered in noise tend to be more confusable than that uttered in a quiet condition.**

B. Frobenius distance of Senior-level / Junior-level speakers

Speakers	Highly-trained speakers		Junior-level speakers	
Environment	Quiet	Noisy	Quiet	Noisy
Frobenius distance	0.11	0.23	0.29	0.63
	0.15		0.42	

- **Highly-trained speakers spoke less confusable emotional speech than junior-level speakers.**

C. Frobenius distance of female and male speakers

Speakers	Female		Male	
Environment	Quiet	Noisy	Quiet	Noisy
Frobenius distance	0.14	0.28	0.22	0.52
	0.20		0.33	

- **Emotional speech uttered by female speakers is less confusable than emotional speech uttered by male speakers.**

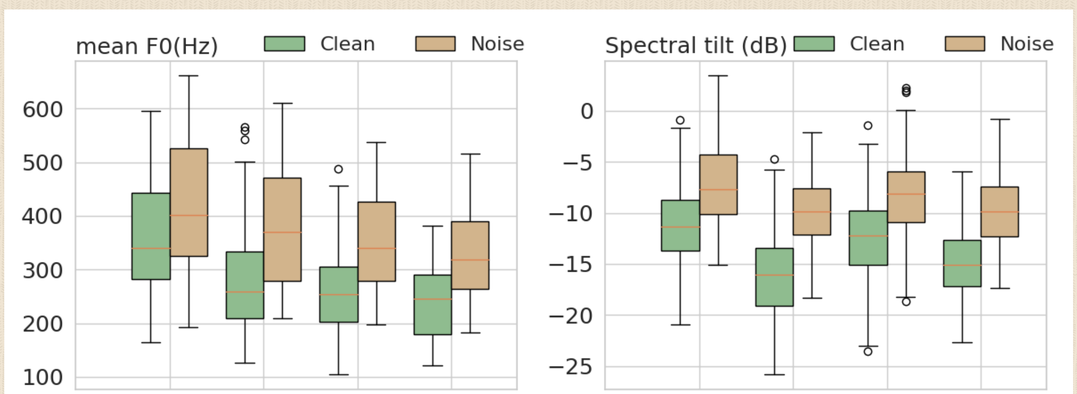
D. Frobenius distance of listeners of different ages

Age of listeners	20s	30s	40s	50s
Frobenius distance	0.17	0.16	0.35	0.50

- **Young listeners can recognize the emotion better than the older listeners.**

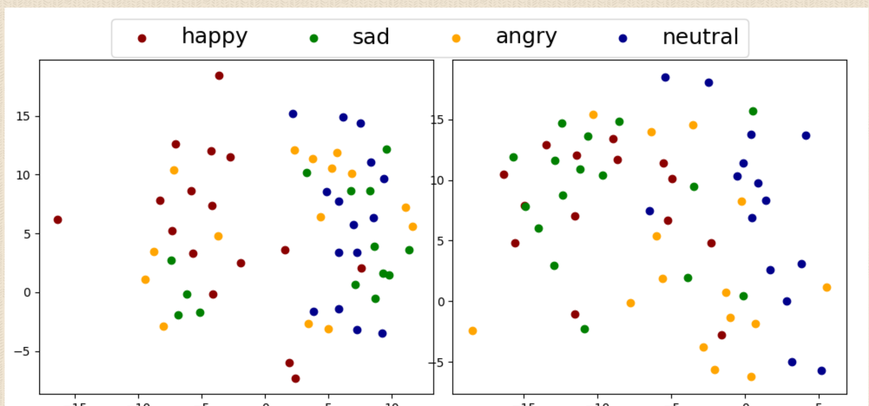
Acoustic Analysis

- Acoustic features of the emotional speech have typical patterns of the Lombard effect.



- **In noisy environment, relative differences among emotional categories become smaller, which makes emotional speech in noise more confusable.**

- We use T-SNE algorithm to visualize acoustic features of one speaker's utterances. Each point represents one utterance.



- **In the quiet condition, most points of sad speech are overlapped with many points of neutral speech. In contrast, in the noisy condition, many sad points are significantly overlapped with those of happy speech.**

Speech Intelligibility

- STOI score: standard objective intelligibility measure and its value ranges between 0 and 1.
- The higher the value is, the more intelligibility the speech is.

Condition	STOI score
Emotional speech produced in noise	0.61
Emotional speech produced in the quiet environment but added the same simulated noise.	0.44

Conclusions

1). If speakers are female or better trained, they can produce less confusable emotions robust to noisy conditions.
2). The recognition accuracy largely depends on the listeners' age.
3). The acoustic differences between emotional speech in quiet and noisy environments depend on the emotion category.
4). Because of interactions with the Lombard effect, relative differences of important acoustic cues among emotion categories become smaller conversely. This is one reason why emotional speech in noise is more easily confused.