21 Turn-taking matters in conversation recognition
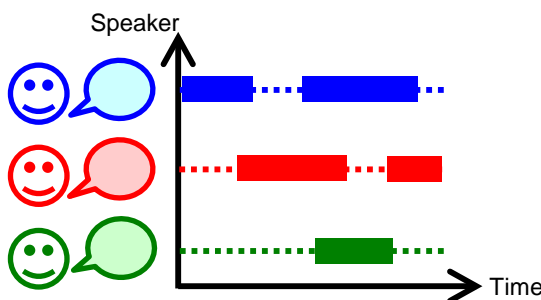
- Robust speech processing using speakers' activity estimation -

Abstract

We present our speech processing technologies developed for conversational speech recognition. Specifically, our focus is on techniques for speaker activity estimation (estimation of each speaker's talking periods), because they play an important role in conversational speech recognition. As shown here, we can enhance a target speech signal from a recorded conversational speech signal by controlling a speech enhancement process according to the estimated speaker activities. It is also possible to improve the speech recognition accuracy by introducing the speaker activity information, including turn-taking information, into the language model in a speech recognition system. Our newly-developed speaker activity estimation method, which is based on a probabilistic model of speaker spatial information, is also presented. With these technologies, we contribute to realizing a more natural voice interface for our daily speech communication.
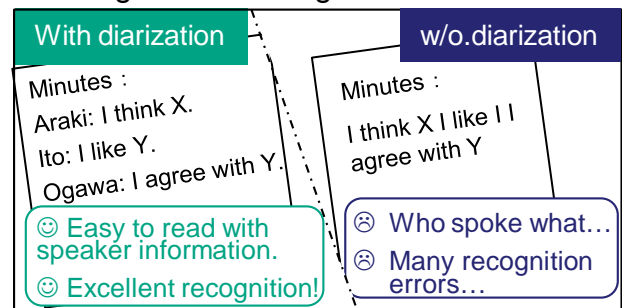
◆Speaker activity estimation:

Technique for estimating the speech periods of each speaker (Also referred to as "diarization")

◆Main applications:
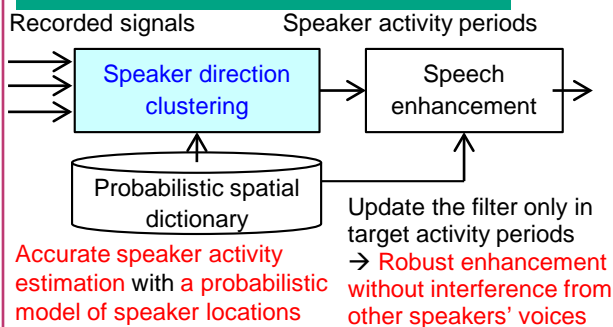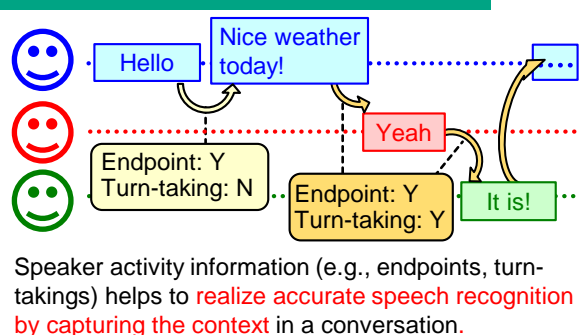
・Meeting minute taking

With diarization

Minutes :
Araki: I think X.
Ito: I like Y.
Ogawa: I agree with Y.

☺ Easy to read with speaker information.
☺ Excellent recognition!

w/o.diarization

Minutes :
I think X I like I I agree with Y

☹ Who spoke what…
☹ Many recognition errors…

・Hearing assistance system

・Communication robot control

◆Speaker activity estimation and its applications to speech processing:

Speaker activity estimation and target speech enhancement

Recorded signals → Speaker direction clustering → Speaker activity periods → Speech enhancement →

Probabilistic spatial dictionary

Update the filter only in target activity periods → Robust enhancement without interference from other speakers' voices

Accurate speaker activity estimation with a probabilistic model of speaker locations

Language model for conversational speech recognition

Hello → Nice weather today! → …

Endpoint: Y Turn-taking: N

Yeah

Endpoint: Y Turn-taking: Y

It is!

Speaker activity information (e.g., endpoints, turn-takings) helps to realize accurate speech recognition by capturing the context in a conversation.

Reference

[1] N. Ito, A. Araki, M. Delcroix, T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Proc. ICASSP2017*, 2017.

[2] H. Ashikawa, T. Morioka, A. Ogawa, T. Iwata, N. Tawara, T. Ogawa, T. Kobayashi, "Recurrent neural network language models using speaker-derived features for multiparty conversations," in *Proc. 2016 Autumn meeting of ASJ*, 3-Q-2, pp. 85-88, Sept. 2016 (in Japanese).

Contact

Shoko Araki Signal Processing Research Group, Media Information Laboratory
Email : araki.shoko(at)lab.ntt.co.jp