Localization Guided Speech Separation

Sunit Sivasankaran
Supervised by: Emmanuel Vincent and Dominique Fohr
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

OVERVIEW

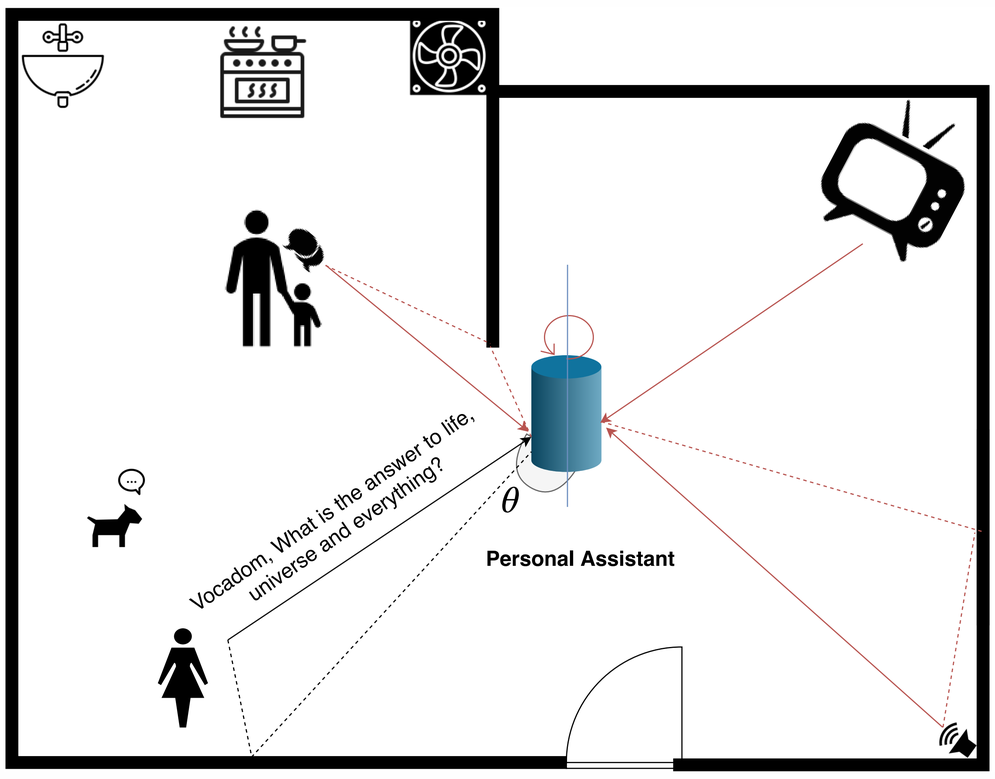
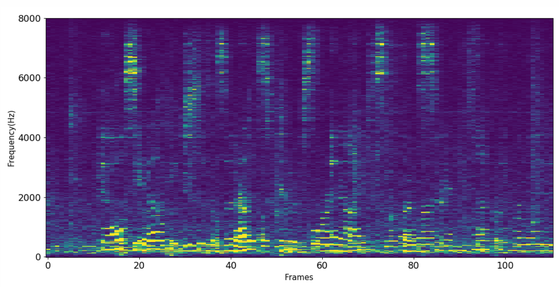
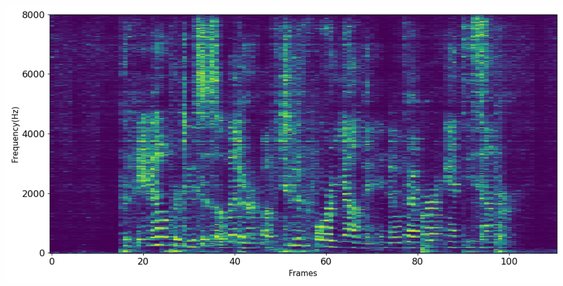


Figure: Sound scene in a typical home



(a) Captured Mixture



(b) Target Source

Figure: Time-Frequency representation of speech

- Mixture contains speech from two speakers and noise
- Typically happens in devices such as Alexa and Google Home
- Interested only in the speaker interacting the device
- Recover speech using speaker location information
- Two step process:
  - Estimate speaker location using known keyword
  - Using the location information to extract the interested speech

SPEAKER LOCALIZATION

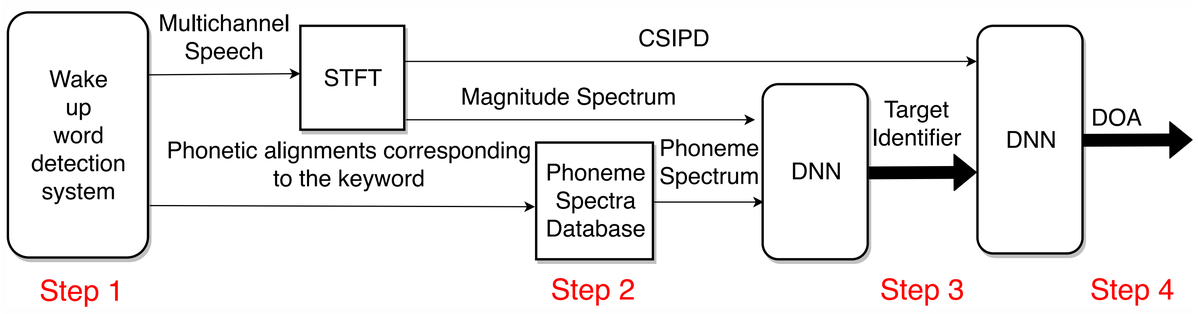
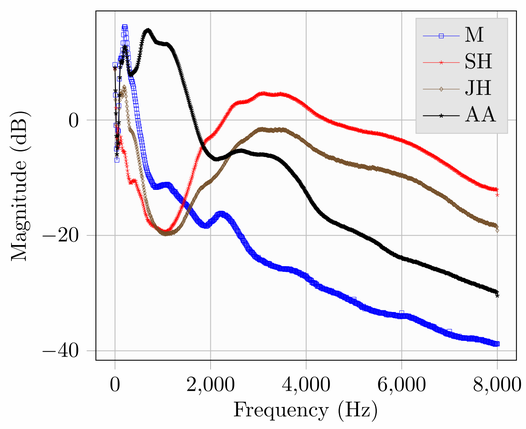


Figure: Flow to estimate the speaker location

- Assumes knowledge of keyword like Alexa, Ok Google
- Words can be broken into phones. Example Alexa: AH L EH K S AH
- Phones has patterns. **Use pattern to improve localization**



SPEECH SEPARATION USING LOCATION INFORMATION

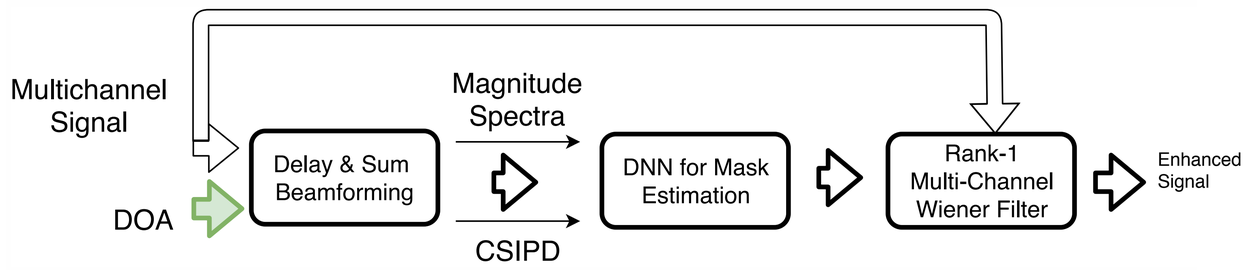


Figure: Speech separation using localization information

- Use the location information to electronically steer towards the speaker
- Extract features to estimate a mask
- Use mask along with beamformers to extract the speaker

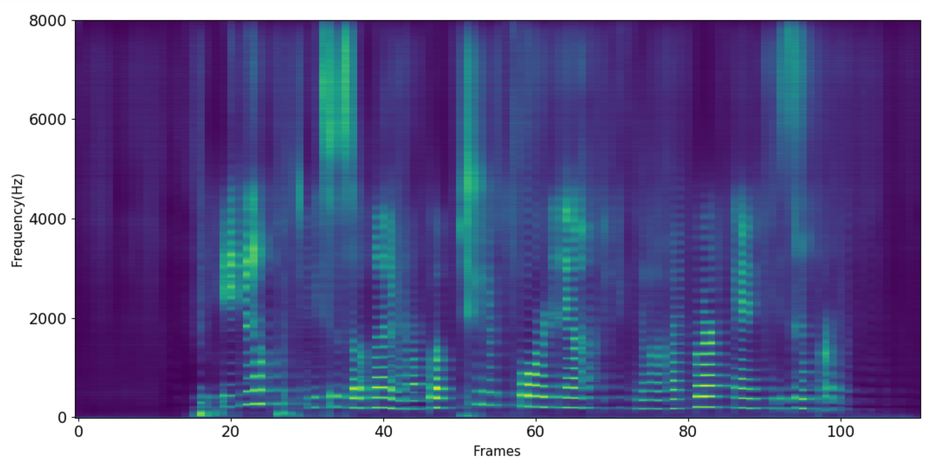


Figure: Estimated speech

RESULTS

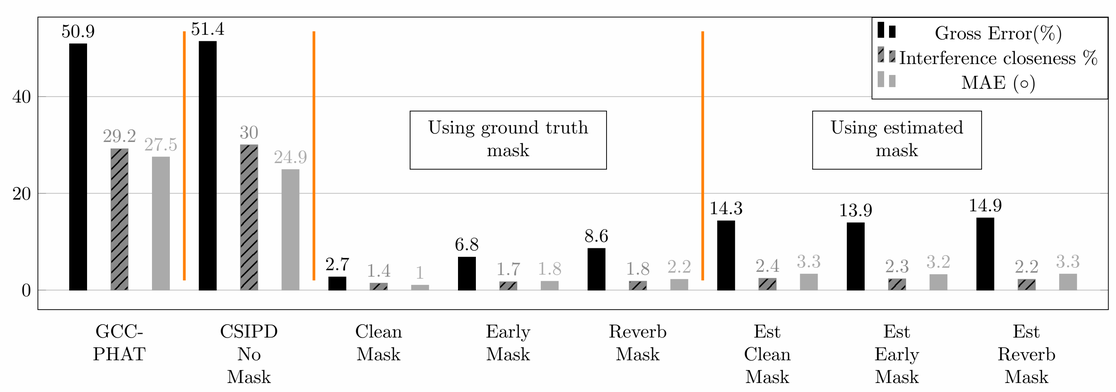


Figure: Localization results

△ DOA	<10°		>50°		**Average**	
Beamformers	**True DOA**	**Est DOA**	**True DOA**	**Est DOA**	**True DOA**	**Est DOA**
GEV	38.0	54.5	30.2	41.5	30.9	43.2
R1-MWF	37.4	53.9	28.8	40.4	29.4	42.4
SDW	36.6	54.0	29.0	40.9	29.6	42.4

Table: Word error rate(%) on noisy two-speaker mixtures after separation using ground truth or estimated speaker location information (Using GCC-PHAT).

CONCLUSION

- Knowledge of text improves localization performance
- Can use localization information to improve separation performance
- Extended this approach to estimate speakers using deflation strategy

PUBLICATIONS

- Sunit Sivasankaran, Emmanuel Vincent, Dominique Fohr, *Keyword-based speaker localization: Localizing a target speaker in a multi-speaker environment*, Interspeech 2018
- Sunit Sivasankaran, Emmanuel Vincent, Dominique Fohr, *Analyzing the impact of speaker localization errors on speech separation for automatic speech recognition* , ICASSP 2020 (submitted)
- Sunit Sivasankaran, Emmanuel Vincent, Dominique Fohr, *SLOGD: Speaker Location Guided Deflation Approach to Speech Separation*, ICASSP 2020 (submitted)

We acknowledge the support of ANR Vocadom and the TALC project.