Keyword based speaker localization

Localizing a target speaker in a multispeaker environment

Sunit Sivasankaran Emmanuel Vincent Dominique Fohr

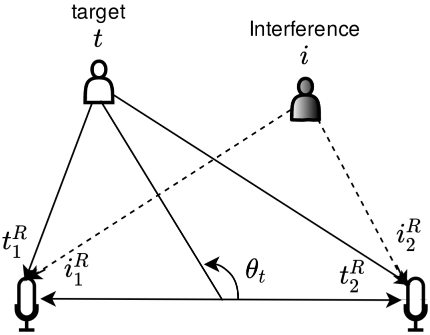
Inria, Nancy, France

LISTEN Workshop, Bonn, Germany
July 17, 2018

Overview

- Prior work on localizing multiple speakers
- Localizing a specific speaker will need further post-processing. Can be error prone
- In our work we localize a speaker who uttered a keyword in a multispeaker environment such as ‘OK Google’ or ‘Alexa’
- New Task
- Two problems :
  - How to identify the intended speaker
  - How to use this identifier information in localization pipeline
- We use time-frequency mask to identify speaker

Problem set up



- Two microphones
- Target and interference speakers speak simultaneously
- Goal is to estimate the DOA of the target θ_t using :
  - The signal $s_c = t_c^R + i_c^R + \eta_c$
  - Keyword (any text) spoken by the target

Learning based approach to localization

- DNN learns a mapping between input features and a discretized DOA space
- Different input features are used :
  - Phasemap = Raw phases of multichannel signals
  - GCC-PHAT features
  - Cosine-Sine Inter channel Phase Difference (CSIPD)
    - A concatenation of cosines and sines of the phase difference between the 2-channel microphones

$$\Delta\phi[\omega, n] = \angle S_1[\omega, n] - \angle S_2[\omega, n] \tag{1}$$

$$\text{CSIPD}[\omega, n] = [\cos(\Delta\phi[\omega, n]), \sin(\Delta\phi[\omega, n])] \tag{2}$$

S_1 and S_2 are STFTs of signals captured at two microphones

Motivation for using CSIPD

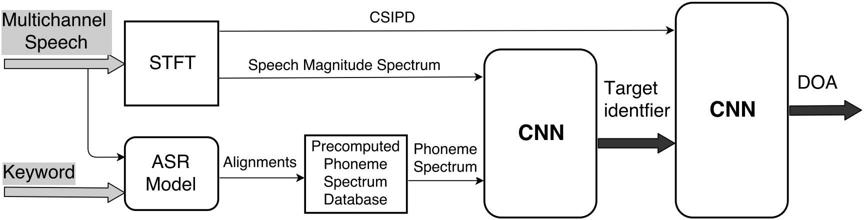
- Motivated by GCC-PHAT

$$C_{1,2}[\tau, n] = \sum_{\omega} \frac{S_1[\omega, n] S_2^*[\omega, n]}{|S_1[\omega, n]||S_2[\omega, n]|} \exp^{j\omega\tau} \quad (3)$$

$$\frac{S_1(\tau, f) S_2^*(\tau, f)}{|S_1(\tau, f)||S_2(\tau, f)|} = \exp^{j\Delta\phi} = \cos(\Delta\phi) + j\sin(\Delta\phi) \quad (4)$$

- Linear projection of CSIPD onto the sinusoidal sub-space $C_{1,2}[\tau, n] = A[\tau, \omega] \times CSIPD[\omega, n]$
- CSIPD with DNN = Non-linear version of GCC-PHAT
- More invariant compared to Phasemap
- Useful to incorporate textual information
- Multiply the mask with CSIPD

Approach



Four step process :

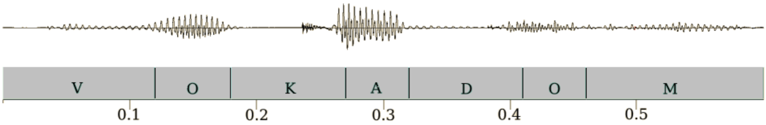
Step1 : Obtain ASR alignments

Step2 : Use alignments to obtain a representative spectra : Phone spectra

Step3 : Estimate target mask and multiply with CSIPD

Step4 : CSIPD $\times$ target mask $\implies$ [DNN] $\implies$ DOA

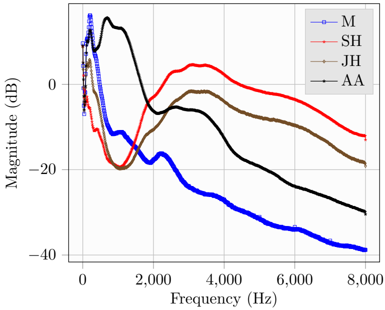
Step1 : ASR alignments



- Wake-up word detects keyword
- Use ASR acoustic model to align speech with text
- HMM-GMM systems used in this work

Step2 : Phone spectra

- Pre-computed by averaging magnitude spectra per phone
- Distinct patterns are observed for every phone
- Pick spectrum corresponding to the aligned phone



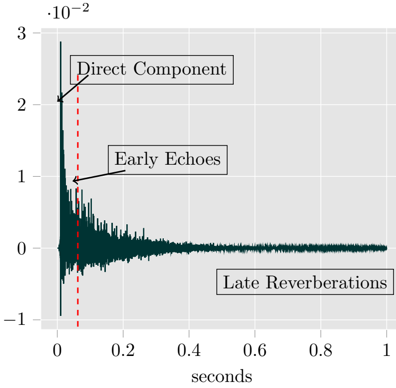
Step3 : Target mask

- Masks represents the amount of target signal in each TF bin
- Three types of mask :
  - Clean target mask, M^D
  - Early target mask, M^E
  - Reverberated target mask

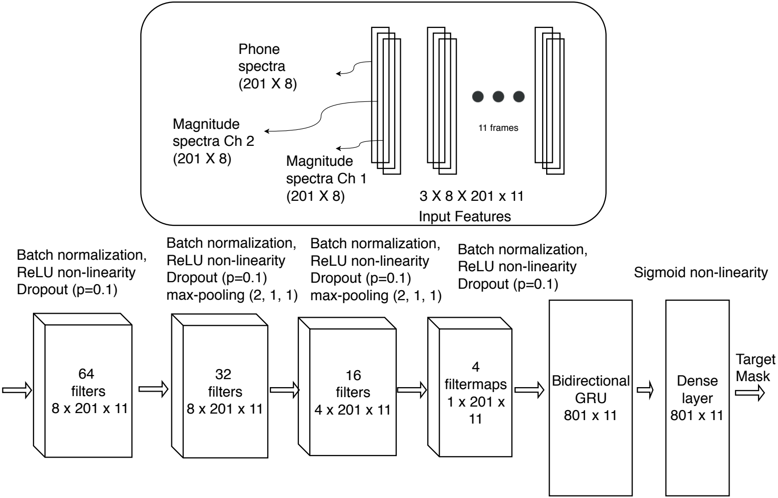
$$\delta_c^E = s_c - t_c^E \tag{5}$$

$$M^E = \frac{|T^E|}{|T^E| + |\Delta^E|} \tag{6}$$

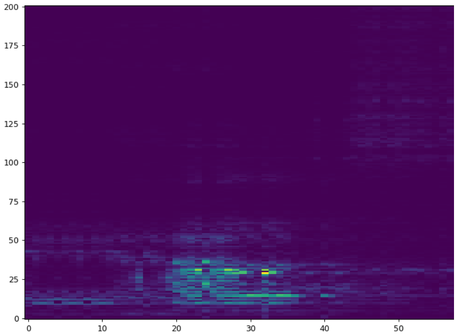
- Need larger frame duration (100 ms) to estimate DOA, but ASR alignments are for short duration (25ms)



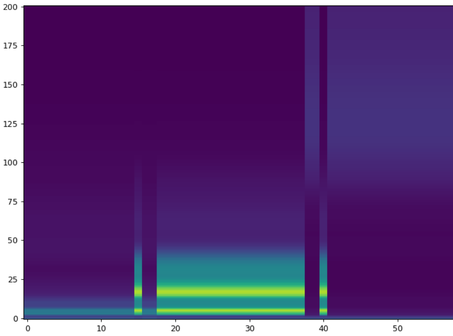
Step3 : Estimating target mask



Step3 : Target mask (contd..)

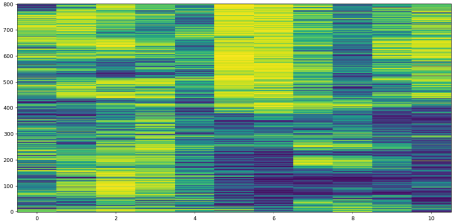


Magnitude Spectrum

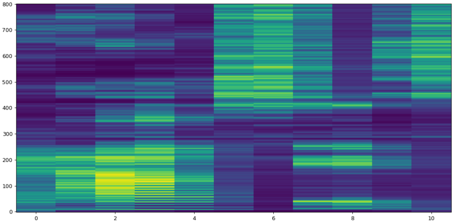


Phone Spectrum

Step3 : Target mask (contd..)

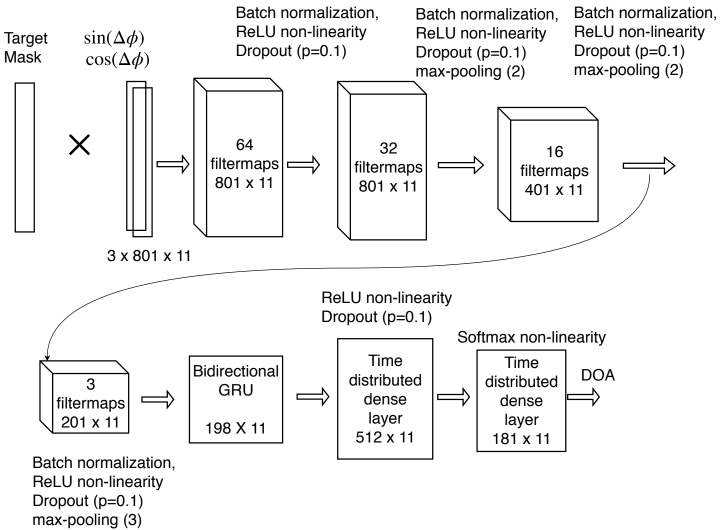


True Mask



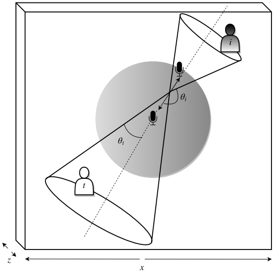
Estimated Mask

Step4 : DOA estimation



Generating RIR

- Discretize DOA space into 1° classes $\implies$ 181 classes
- Create all possible target DOA and interference DOA pairs $\{\theta_t, \theta_i\}, \forall\theta_t \in [0, 180], \forall\theta_i \in [0, 180]$ with the constraint $|\theta_t - \theta_i| > 5^\circ$
- 50,1 and 2 such positions are created for every θ_i, θ_j for training, validation and test
- This resulted in 1557600, 31152, and 62304 configurations
- RIR simulated using RIR-Simulator



Feature extraction

- Speech signals from Librispeech
- Two 0.5 s segments are randomly picked and convolved with the target and inference RIRs from a single room
- Signal-to-Interference ratio (SIR) $[0, 10]$ dB
- Speech shaped noise (SSN) for training at SNR $[0, 15]$ dB
- Real ambient noise for test at SNR $[0, 30]$ dB

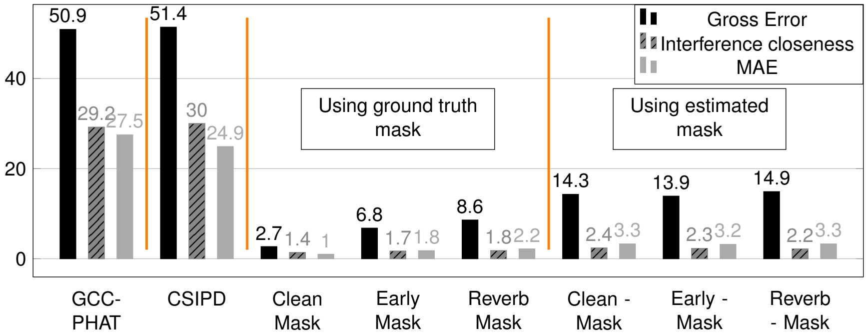
Metrics

Gross Error rate : % of estimated DOAs above a 5° error tolerance

Interference closeness rate : % of estimated DOAs which are close ($< 5^\circ$) to the interference DOA

Mean absolute error(MAE) : Mean of the absolute error with respect to Target DOA (in degrees)

Results



- Target mask helps in identifying the target
- Estimated mask has low interference closeness rate

With noisy alignments

	Clean Mask		Early Mask		Reverb Mask	
Alignments	Noisy	Clean	Noisy	Clean	Noisy	Clean
Gross Error Rate	15.2	14.3	14.8	13.9	15.9	14.9
Interference Close Rate	2.5	2.4	2.4	2.3	2.3	2.2
MEA	3.8	3.3	3.6	3.2	3.9	3.3

Using noisy alignments has negligible effect on performance

Other observations

- Other target identifiers : Spectrum based
- Mask identifiers works better than spectrum identifier
- Multiplying masks with CSIPD is better than appending
- Fricatives are better suited for localization and nasal are the worst

Phone	CH_I	CH_B	Z_B	SH_B	NG_E	N_E	M_E	B_B
Error rate	1.5	1.6	1.8	1.8	19.4	21.1	21.3	24.5

Conclusion

- Proposed methods to incorporate text into speaker localization pipeline
- Masks are good target identifiers. Multiply $>$ Append
- Fricatives phones are better for localization and plosive sounds are the worst
- Ok Google sshhhhhhhhhhhhhhh ! ! !

Thank you