A Multichannel Feature Compensation Approach for Robust ASR in Noisy and Reverberant Environments

Ramón F. Astudillo [1] Sebastian Braun [2] Emanuël A. P. Habets [2]

[1]Spoken Language Systems Laboratory, INESC-ID-Lisboa
Lisboa, Portugal

[2]International Audio Laboratories Erlangen
Am Wolfsmantel 33, 91058 Erlangen, Germany

Overview of the Proposed System

The approach integrates STFT-domain enhancement with the ASR system through Uncertainty Propagation.

Three main components detailed:

- Joint reverberation and noise reduction by informed spatial filtering applied in STFT domain.
- Multichannel MMSE-MFCC estimator with different STFT configurations for enhancement and recognition domains.
- Model-based feature enhancement using the MSE of the MMSE-MFCC estimator and Modified Imputation.

Joint reverberation and noise reduction

- Signal model: single source $S(k,n)$, propagation vector $\mathbf{d}(k,n)$, reverberation $\mathbf{r}(k,n)$ and additive noise $\mathbf{v}(k,n)$

$$\mathbf{y}(k,n) = \mathbf{d}(k,n)S(k,n) + \mathbf{r}(k,n) + \mathbf{v}(k,n)$$

- All components mutually uncorrelated with variances equal to

$$\begin{aligned}\mathbf{\Phi}_{\mathbf{y}}(k,n) = \phi_S(k,n)\,\mathbf{d}(k,n)\mathbf{d}^H(k,n) + \phi_R(k,n)\,\mathbf{\Gamma}_{\text{diff}}(k) \\ + \mathbf{\Phi}_{\mathbf{v}}(k,n)\end{aligned}$$

- Multichannel minimum MSE (M-MMSE) source estimate:

$$\begin{aligned}\hat{S}_{\text{M-MMSE}}(k,n) &= \underset{\hat{S}(k,n)}{\arg\min}\, E\left\{|S(k,n) - \hat{S}(k,n)|^2\right\} \\ &= \underbrace{H_{\text{MMSE}}(k,n) \cdot \mathbf{h}_{\text{MVDR}}(k,n)}_{\mathbf{h}_{\text{M-MMSE}}(k,n)}{}^H\, \mathbf{y}(k,n)\end{aligned}$$

Joint reverberation and noise reduction

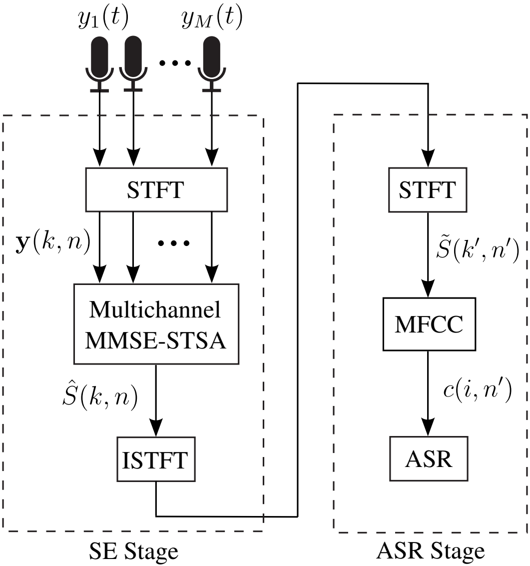
Optional use of multichannel MMSE Amplitude (M-STSA) estimate:

$$\hat{S}_{\text{M-STSA}}(k,n) = \underbrace{H_{\text{STSA}}(k,n) \cdot \mathbf{h}_{\text{MVDR}}(k,n)}_{\mathbf{h}_{\text{M-STSA}}(k,n)}{}^{H}\, \mathbf{y}(k,n)$$

Parameter estimation per time-frequency

- DOA for $\mathbf{d}(k,n)$: Beamspace root-MUSIC (circular array) [Zoltowski et al. 1992]
- Diffuse PSD $\phi_R(k,n)$: maximum likelihood estimator [Braun 2013 et al.]
- Noise covariance matrix $\mathbf{\Phi}_{\mathbf{v}}(k,n)$: speech presence probability based recursive estimation [Souden 2011 et al.]

Joint reverberation and noise reduction



M-MMSE-MFCC estimator

In the context of ASR, MMSE-MFCC estimators [Yu 2008], [Astudillo 2010], [Stark 2011], bring interesting advantages

- Same signal model as STFT domain estimators e.g. Wiener, MMSE-STSA, MMSE-LSA.
- The approach in [Astudillo 2010], here used, also provides the minimum MSE in MFCC domain.
- The same approach can be applied to derive a M-MMSE-MFCC estimator from the M-MMSE

M-MMSE MFCC Estimator

The posterior distribution for the M-MMSE is given by

$$p(S(k,n)|\mathbf{y}(k,n)) \sim \mathcal{N}_{\mathbb{C}}\left(\hat{S}_{\text{M-MMSE}}(k,n), \lambda(k,n)\right),$$

where the variance is equal to the minimum MSE

$$\begin{aligned}\lambda(k,n) &= E\left\{|S(k,n) - \hat{S}_{\text{M-MMSE}}(k,n)|^2\right\} \\ &= \phi_S(k,n)(1 - \mathbf{h}^H_{\text{M-MMSE}}(k,n)\mathbf{d}(k))\end{aligned}$$

In theory, the posterior for the M-MMSE-MFCC can be obtained by Uncertainty Propagation as

$$p(c(i,n)|\mathbf{y}(n)) \sim \mathcal{N}_{\mathbb{C}}\left(\hat{c}^{\text{M-MMSE-MFCC}}(i,n), \lambda^c(i,n)\right).$$

M-MMSE MFCC Estimator

In practice, we need to propagate variances through the STFT.

Let $\phi(n)$ be the variance of speech or noise, the variance after ISTFT+STFT is given by

$$\tilde{\phi}(n') = \sum_{n \in \mathrm{Ov}(n')} |\mathbf{R}_{n'-n}|^2 \phi(n),$$

- $\mathbf{R}_{n'-n}$ is built by multiplying the inverse Fourier and Fourier matrices truncated to the corresponding overlap.
- Summing over all overlapping frames Ov attenuates variance artifacts (STFT consistency).
- Correlations induced by overlapping windows ignored.

Model-based feature enhancement

Since the minimum MSE of the M-MMSE-MFCC is available we can apply observation uncertainty techniques.

Modified Imputation [Kolossa 2005] showed the best performance, this is given by

$$\hat{c}_q^{\mathrm{MI}}(i,n') = \frac{\Sigma_q(i)}{\Sigma_q(i) + \lambda^c(i,n')} \hat{c}^{\mathrm{M\text{-}MMSE}}(i,n') + \frac{\lambda^c(i,n')}{\Sigma_q(i) + \lambda^c(i,n')} \, \mu_q(i), \tag{1}$$

where $\boldsymbol{\mu}_q$ and $\boldsymbol{\Sigma}_q$ are the mean and variances of the q-th ASR Gaussian mixture.

Proposed System

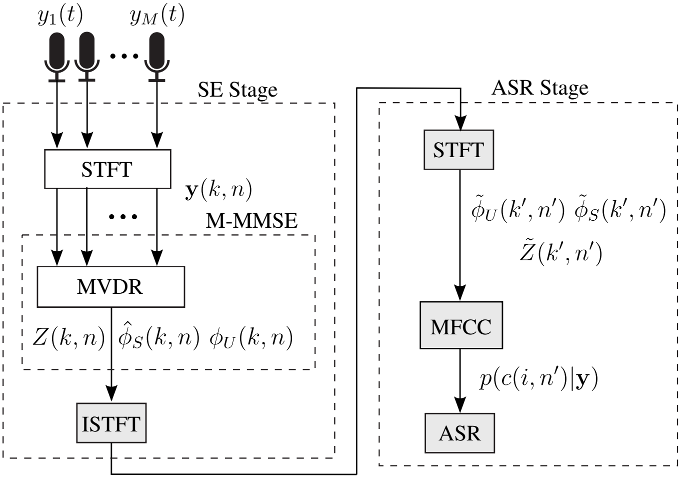
Characteristics

- M-MMSE-MFCC with optional use of MI as described.
- System is real-time capable, per-frame batch if CMS used.
- To improve performance, speech variance $\phi_S(k, n)$ re-estimated using the M-STSA.

Implementation

- M-STSA, M-MMSE-MFCC implemented in Matlab.
- Modified version of HTK used for MI.

Proposed System



Beamformed signal: $Z(k,n) = \mathbf{h}_{\mathrm{MVDR}}(k,n)^H \mathbf{y}(k,n)$

Residual variance: $\phi_U(k,n) = \mathbf{h}_{\mathrm{MVDR}}^H (\phi_R \, \mathbf{\Gamma}_{\mathrm{diff}} + \mathbf{\Phi}_{\mathbf{v}}) \mathbf{h}_{\mathrm{MVDR}}$

REVERB 2014 Results

HTK baseline, development set results for clean training

Simulated Data							
	Room 1		Room 2		Room 2		Avg.
	Near	Far	Near	Far	Near	Far	
No Proc.	**14.43**	25.15	43.46	86.64	52.20	88.40	51.67
MSTSA	19.25	27.65	18.68	36.55	24.60	47.16	28.97
M-MFCC	16.94	23.57	17.20	**33.47**	**20.80**	**44.29**	26.03
+MI	15.34	**21.85**	**16.96**	33.67	20.99	45.03	**25.64**

Recorded Data			
	Room 1		Avg.
	Near	Far	
No Proc.	88.33	87.56	87.94
MSTSA	58.27	61.18	59.71
M-MFCC	54.15	54.41	54.27
+MI	**51.72**	**50.31**	**51.02**

REVERB 2014 Results

HTK baseline, development set results for multi-condition training

Simulated Data							
	Room 1		Room 2		Room 2		Avg.
	Near	Far	Near	Far	Near	Far	
No Proc.	16.54	18.88	23.37	43.18	27.40	46.79	29.34
MSTSA	15.46	17.75	17.23	26.13	18.40	30.91	20.97
M-MFCC	15.73	16.79	14.81	**21.99**	18.05	**27.35**	19.11
+MI	**14.70**	**16.74**	**14.30**	23.05	**17.80**	27.42	**19.00**

Recorded Data			
	Room 1		Avg.
	Near	Far	
No Proc.	52.90	50.79	51.85
MSTSA	42.48	41.49	41.98
M-MFCC	40.61	39.23	39.92
+MI	**39.74**	**37.18**	**38.46**

Conclusions

- Improvements over M-STSA by integration with ASR.
- Results for real data worse compared to simulated data, but consistent across methods.
- The use of observation uncertainty (MI) yields good results in highly mismatched situations.
- ISTFT+STFT propagation simplifies integration with well established STFT-domain methods.

Thank You!

MMSE-MFCC Matlab code available under

https://github.com/ramon-astudillo/stft_up_tools

MI HTK patches available under

http://www.astudillo.com/ramon/research/stft-up/