USE OF MULTIPLE FRONT-ENDS AND I-VECTOR-BASED SPEAKER ADAPTATION FOR ROBUST SPEECH RECOGNITION

Md Jahangir Alam[1,2], Vishwa Gupta[1], Patrick Kenny[1], Pierre Dumouchel [2]

[1] CRIM, Montreal, Canada, [2] École de technologie supérieure, Montréal, Canada

ÉTS Le génie pour l'industrie

CRIM

Summary

✓ Although state-of-the-art speech recognition systems perform well in controlled environments they work poorly in realistic acoustical conditions in reverberant environments.

✓ We use multiple front-ends - based recognition systems with multi-condition training data and combine their results using ROVER (Recognizer Output Voting Error Reduction).

✓ For 2- and 8- channel tasks, to get benefit from more than one channel, we utilize ROVER instead of the multi-microphone signal processing method.

✓ As in previous work we also apply i-vector-based speaker adaptation which was found effective.

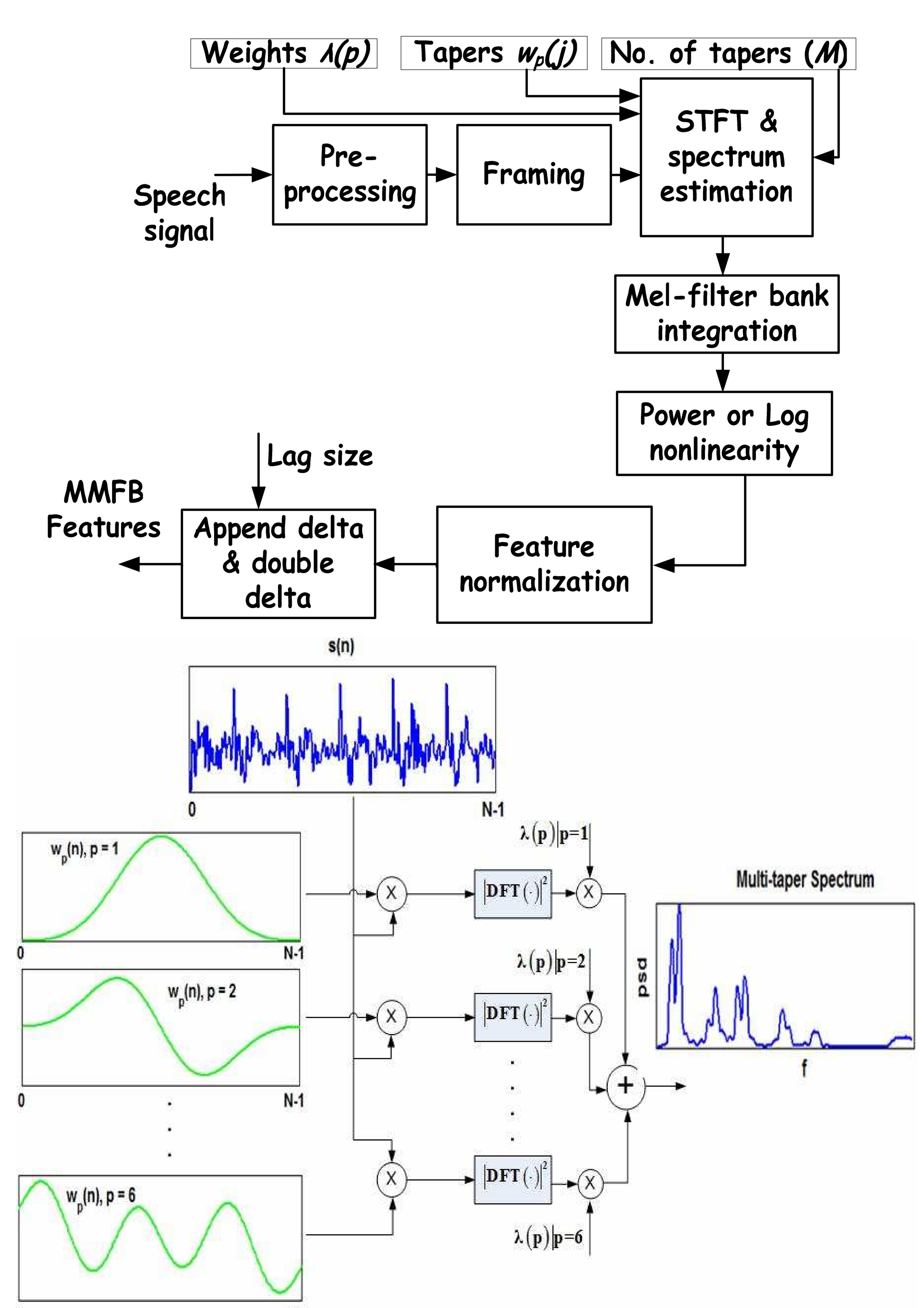
✓ Speech recognition experiments using the DNN-HMM hybrid architecture are conducted on the REVERB challenge 2014 corpora using the Kaldi recognizer.

✓ For the 2-channel task (using full batch processing) we obtained an average word error rate (WER) of 9.0% and 23.4% on the SimData and RealData respectively. Whereas for 8-channel task on the SimData and RealData the average WERs found were 8.9% and 21.7%, respectively.

Front-Ends

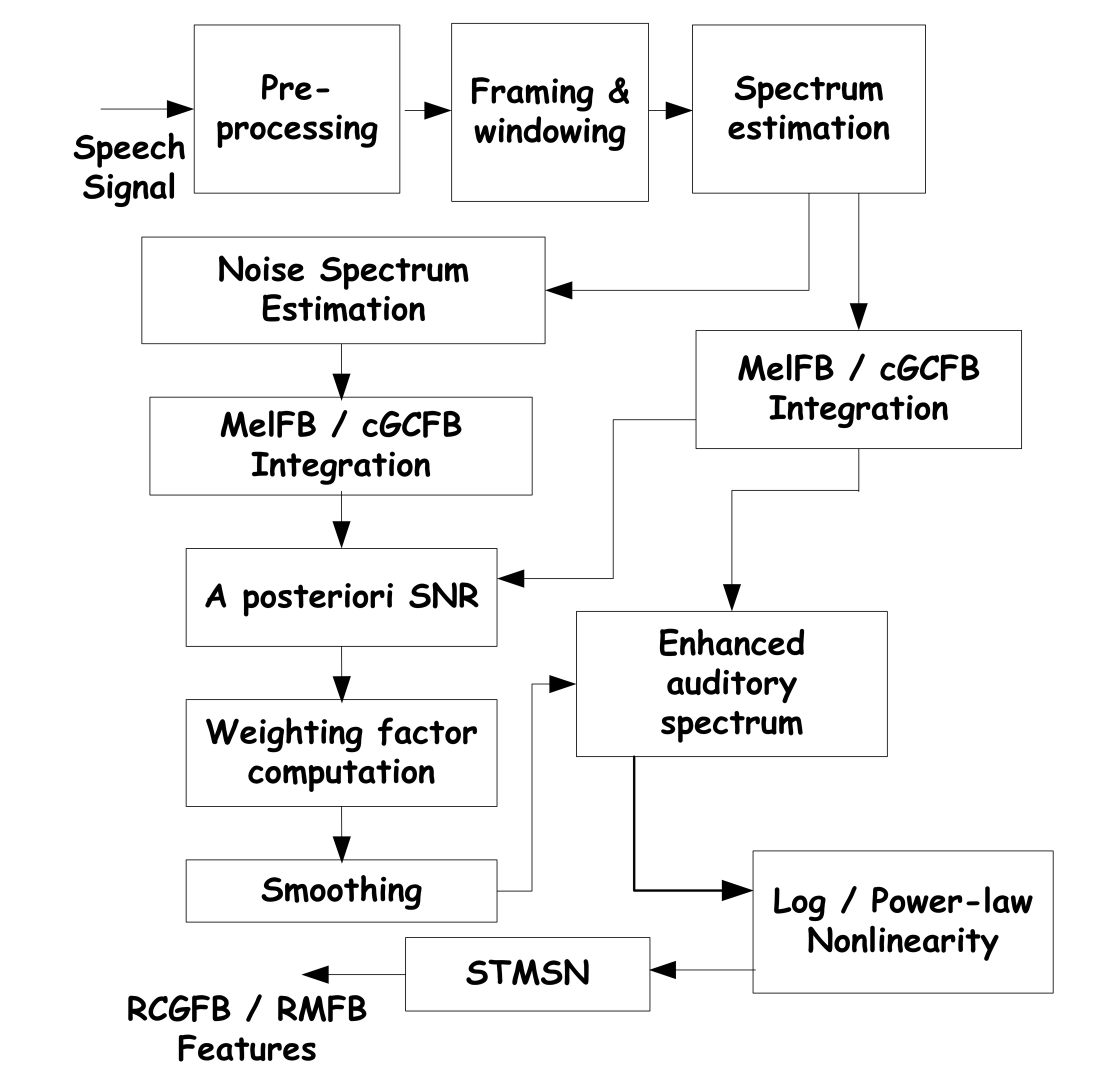
The following front-ends (or feature extractors) were considered in this REVERB Challenge 2014 task.

Multitaper Mel-Filterbank (MMFB) Features

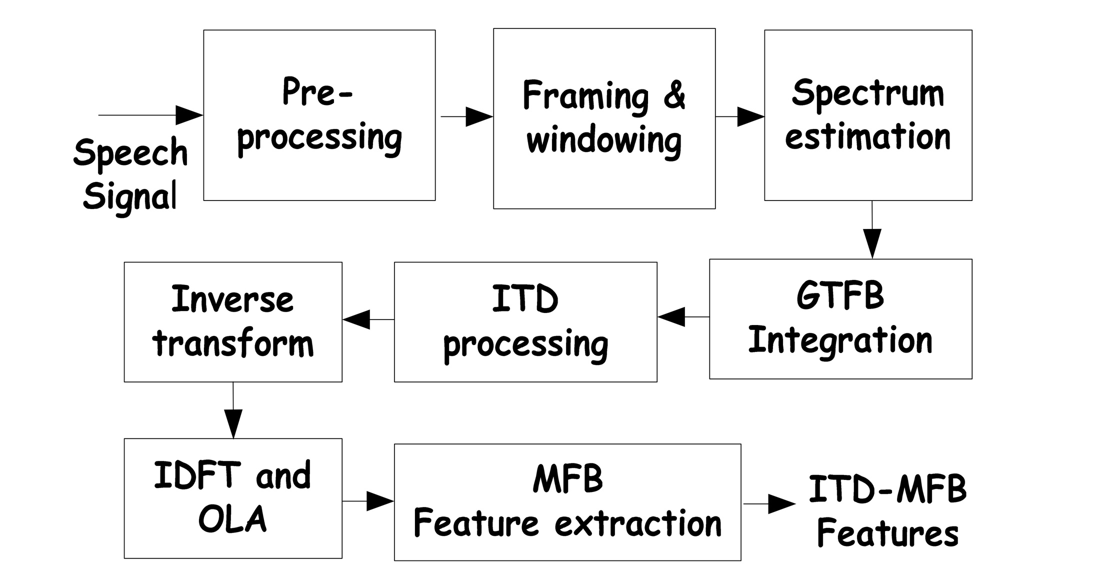


$$\mathrm{var}\left(\hat{S}_{MT}(m,k)\right) \approx \frac{1}{M}\mathrm{var}\left(\hat{S}_d(m,k)\right)$$

Robust Compressive Gammachirp Filterbank (RCGFB) & Robust Mel-Filterbank (RMFB) Features

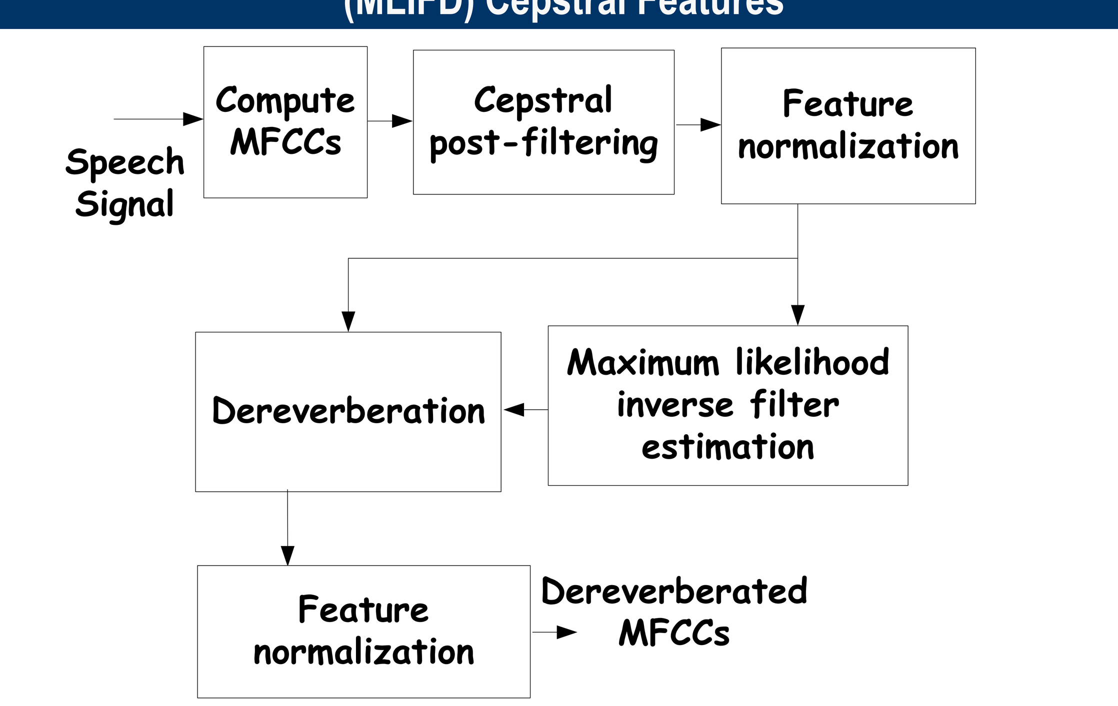


Iterative Deconvolution (ITD)-based MFB (ITD-MFB) Features



MFB Feature extraction steps are similar to the MMFB feature extraction. In MFB, for spectrum estimation, Hamming windowed periodogram is used instead of the multitaper spectrum estimator.

Maximum Likelihood Inverse Filtering-based Dereverberated (MLIFD) Cepstral Features



$P(z) = 1/\left(1 + pz^{-1}\right) \rightarrow$ IIR dereverberated filter of M taps

$c^d[m] \rightarrow$ dereverberated cepstral features of m-th frame

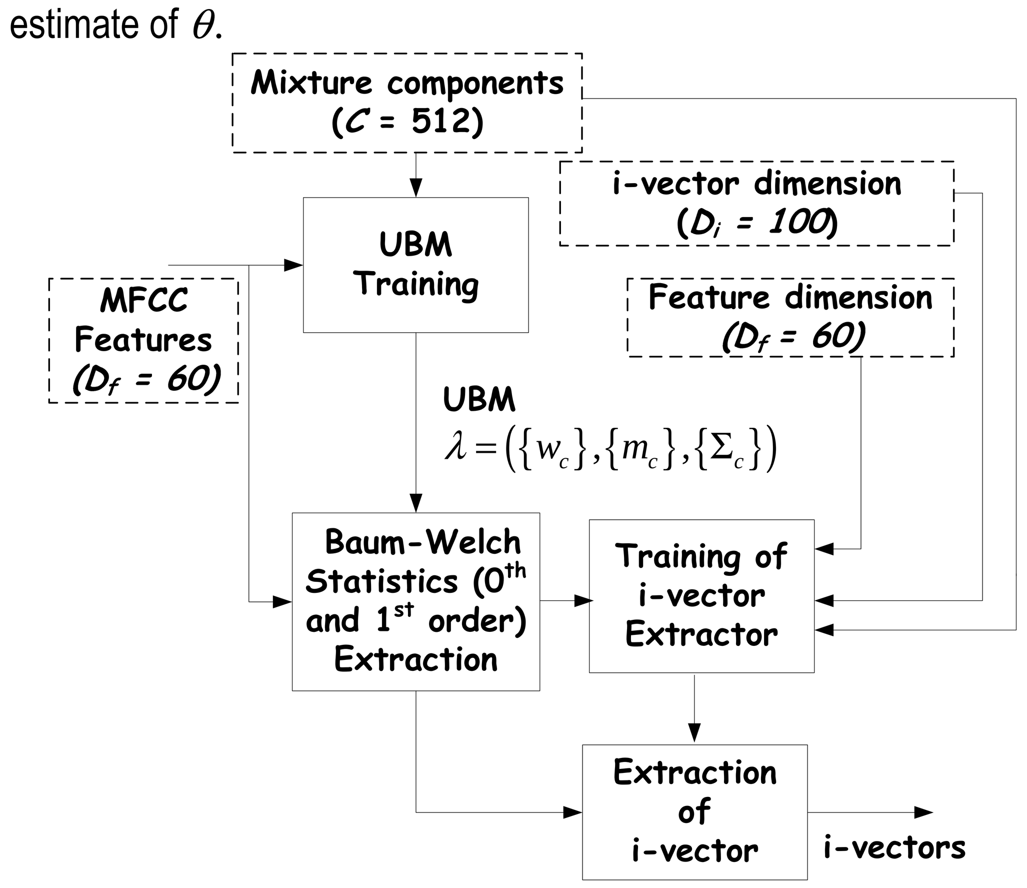
$$c^d[m] = c[m] - \sum_{k=1}^{M-1} p[k] c^d[m-k]$$

Extraction of I-Vectors for Speaker Adaptation

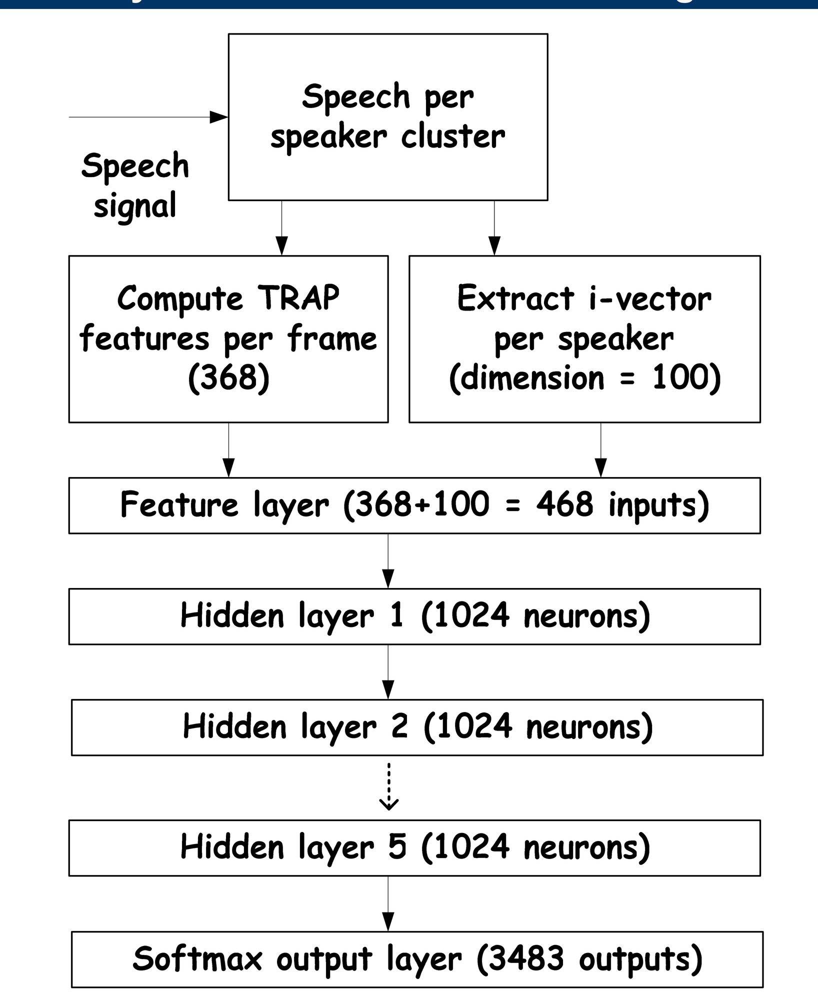
$$M = \mu_c + \mathbf{T}\theta$$

$M \rightarrow$ supervector (speaker & channel dependent), $\mathbf{T} \rightarrow$ Total variability matrix, $\mu_c \rightarrow$ UBM supervector, $c \rightarrow$ mixture component

For each speech recording r, an i-vector i_r is obtained as the MAP estimate of θ.



DNN-HMM Hybrid Architecture for Training & Decoding



✓ Multiple recognition experiments with multiple feature extractors.

✓ Combination of decoded transcripts using ROVER, a recognition system combination software available from NIST.

✓ Baseline system utilizes conventional MFB features.

✓ Every recognition system uses a Deep Neural Networks-HMM (DNN-HMM) hybrid architecture.

✓ For all DNN's a 100-dimensional i-vector is also used with the TRAP features. The i-vector characterizing a speaker helps the DNN to adapt to the speaker characteristics.

✓ For decoding a pruned trigram LM with 709K trigrams generated from the WSJ LM training data.

✓ The resulting lattices were rescored using a larger trigram LM with 3,15 million trigrams generated from the WSJ LM training data.

✓ A vocabulary of 20K words is used.

Experimental Results on the Eval data

Table 1 *WER obtained using utterance-based batch processing for (a) 1-, (b) 2-, & (c) 8-channel tasks.*

(a)	SimData							RealData		
	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
	Near	Far	Near	Far	Near	Far		Near	Far	
RCGFB	8.2	9.4	9.8	15	10.8	16.6	11.6	30.4	31.5	30.9
$MMFB_l$	8.3	9.3	9.9	15.7	10.6	17.6	11.9	31.6	30.9	31.2
$MMFB_p$	8.5	9.8	11.1	17.2	11.8	19.2	12.9	30.2	31.9	31
RMFB	8.4	9.1	9.7	15	10.8	17	11.6	31.8	31.3	31.5
ITD-MFB	7.6	8.8	10.4	14.5	9.8	16	11.1	31.9	33	32.4
Baseline	7.6	8.9	11.5	18.1	11.2	18.8	12.6	41	38	39.4
ROVER-all	**7.1**	**8.1**	**8.9**	**12.9**	**9.2**	**13.8**	**10**	**27.2**	**26.9**	**27.1**

(b)	SimData							RealData		
	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
	Near	Far	Near	Far	Near	Far		Near	Far	
RCGFB	8.4	9.5	10.1	15.2	11.1	17.1	11.9	31.4	32.4	31.9
$MMFB_l$	8.5	9.3	10.1	15.9	11.3	17.9	12.2	32.8	31.2	32
$MMFB_p$	8.9	10	11	18.1	12.5	20.3	13.5	31.8	32.9	32.3
RMFB	8.4	9.1	10	15.4	10.8	17.3	11.9	32.6	31	31.8
ITD-MFB	7.8	9	10.5	15.1	10.4	16.1	11.5	33	32.6	32.8
Baseline	7.8	9.2	11.6	18.3	11.7	19.3	13	42.4	38.5	40.4
ROVER-all	**7**	**7.8**	**8.4**	**12.1**	**9**	**13.2**	**9.6**	**25.5**	**25.7**	**25.6**

(c)	SimData							RealData		
	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
	Near	Far	Near	Far	Near	Far		Near	Far	
RCGFB	8.1	9	9.1	14	10.3	15.3	11	27.9	28.7	28.3
$MMFB_l$	8.1	8.7	9.3	14.3	10.1	15.8	11.1	28.4	27	27.7
$MMFB_p$	8.4	9.2	10.2	16.1	11.4	18	12.2	27.5	28.7	28.1
RMFB	8.1	8.7	9.1	13.6	10	14.8	10.7	29.4	27.7	28.5
ITD-MFB	7.2	8.1	9.7	13.1	9.5	14.8	10.4	29.8	30.1	30
Baseline	7.6	8.4	10.6	17	10.6	17.9	12	37.8	36.8	37.3
ROVER-all	**6.7**	**7.3**	**8.3**	**11.6**	**8.6**	**12.6**	**9.1**	**23.8**	**24.1**	**24**

Table 2 *WER obtained using full batch processing for (a) 1-, (b) 2-, & (c) 8-channel tasks.*

(a)	SimData							RealData		
	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
	Near	Far	Near	Far	Near	Far		Near	Far	
MLIFD	8	8.8	10.9	15.9	11.1	17.6	12	27	28	27.5
RCGFB	8.2	9.4	9.8	15	10.8	16.6	11.6	30.4	31.5	30.9
$MMFB_l$	8.7	9.9	10.3	17.2	11.3	18.7	12.6	28.7	28.7	28.6
$MMFB_p$	8.5	9.8	11.1	17.2	11.8	19.2	12.9	30.2	31.9	31
RMFB	8.4	9.1	9.7	15	10.8	17	11.6	31.8	31.3	31.5
ITD-MFB	7.6	8.8	10.4	14.5	9.8	16	11.1	31.9	33	32.4
Baseline	7.6	8.9	11.5	18.1	11.2	18.8	12.6	41	38	39.5
ROVER-all	**6.7**	**7.3**	**8.4**	**11.8**	**8.7**	**12.7**	**9.3**	**23.8**	**24.8**	**24.3**

(b)	SimData							RealData		
	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
	Near	Far	Near	Far	Near	Far		Near	Far	
MLIFD	8.2	9	11.1	16.5	11.5	18.4	12.5	27.3	27.5	27.4
RCGFB	8.4	9.5	10.1	15.2	11.1	17.1	11.9	31.4	32.4	31.9
$MMFB_l$	8.8	10.4	10.5	17.9	11.8	19.5	13.2	29.5	28.8	29.1
$MMFB_p$	8.9	10	11	18.1	12.5	20.3	13.5	31.8	32.9	32.3
RMFB	8.4	9.1	10	15.4	10.8	17.3	11.9	32.6	31	31.8
ITD-MFB	7.8	9	10.5	15.1	10.4	16.1	11.5	33	32.6	31
Baseline	7.8	9.2	11.6	18.3	11.7	19.3	13	42.4	33	40.4
ROVER-all	**6.6**	**7.4**	**8.1**	**11.2**	**8.5**	**12.2**	**9**	**22.6**	**24.2**	**23.4**

(c)	SimData							RealData		
	Room 1		Room 2		Room 3		Avg.	Room 1		Avg.
	Near	Far	Near	Far	Near	Far		Near	Far	
MLIFD	7.5	8.3	10	14.1	10.4	15.9	11	23.8	24.4	24.1
RCGFB	8.1	9	9.1	14	10.3	15.3	11	27.9	28.7	28.3
$MMFB_l$	8.5	9.5	9.5	16.1	10.8	17.4	12	26	26.2	26.1
$MMFB_p$	8.4	9.2	10.2	16.1	11.4	18	12.2	27.5	28.7	28.1
RMFB	8.1	8.7	9.1	13.6	10	14.8	10.7	29.4	27.7	28.5
ITD-MFB	7.2	8.1	9.7	13.1	9.5	14.8	10.4	29.8	30.1	30
Baseline	7.6	8.4	10.6	17	10.6	17.9	12	37.8	36.8	37.3
ROVER-all	**6.7**	**7.3**	**8**	**11.1**	**8.1**	**12.1**	**8.9**	**21.4**	**22**	**21.7**

Acronyms	
RCGFB	Robust Compressive Gammachirp FilterBank
RMFB	Robust Mel FilterBank
$MMFB_p$	Multitaper Mel FilterBank with power-law nonlinearity
$MMFB_l$	Multitaper Mel FilterBank with logarithmic nonlinearity
ITD-MFB	ITerative Deconvolution-based Mel FilterBank
MLIFD	Maximum Likelihood Inverse Filtering-based Dereverberated cepstrum
MFB	Mel FilterBank (Baseline)
ROVER	Recognizer Output Voting Error Reduction
STMSN	Short-Term Mean and Scale Normalization
UBM	Universal Background Model
DNN	Deep Neural Networks
HMM	Hidden Markov Model
TRAP	TempoRAl Pattern
GTFB/cGCFB	GammaTone/compressive Gammachirp FilterBank
MFCC	Mel-Frequency Cepstral Coefficients