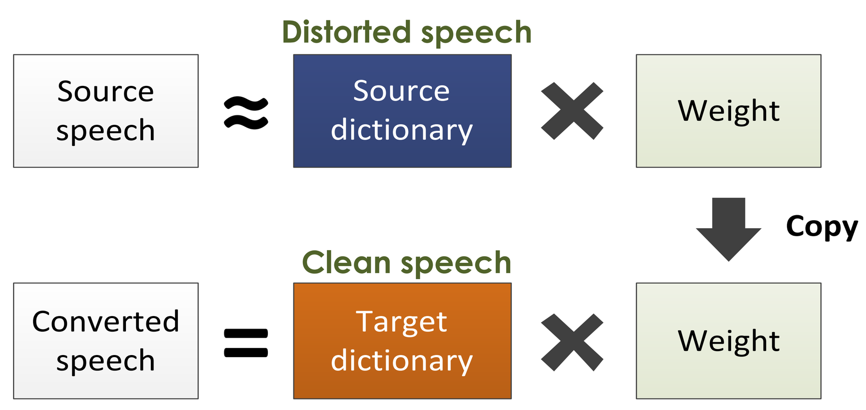
A POST-FILTERING APPROACH BASED ON LOCALLY LINEAR EMBEDDING DIFFERENCE COMPENSATION FOR SPEECH ENHANCEMENT

Yi-Chiao Wu, Hsin-Te Hwang, Syu-Siang Wang, Chin-Cheng Hsu, Yu Tsao, and Hsin-Min Wang
Academia Sinica, Taiwan

Previous work



- Voice conversion : given source speech, letting machine generate target speech according to the relationship between source and target training speech
- Exemplar-based voice conversion : using matrix factorization and source-target paired dictionary (formed by speech features) to convert speech
- (ICASSP2017) Directly using Locally Linear Embedding exemplar-based voice conversion to convert enhanced (distorted) speech to clean speech : post-filtering to compensate the distortion caused by speech enhancement

Problem

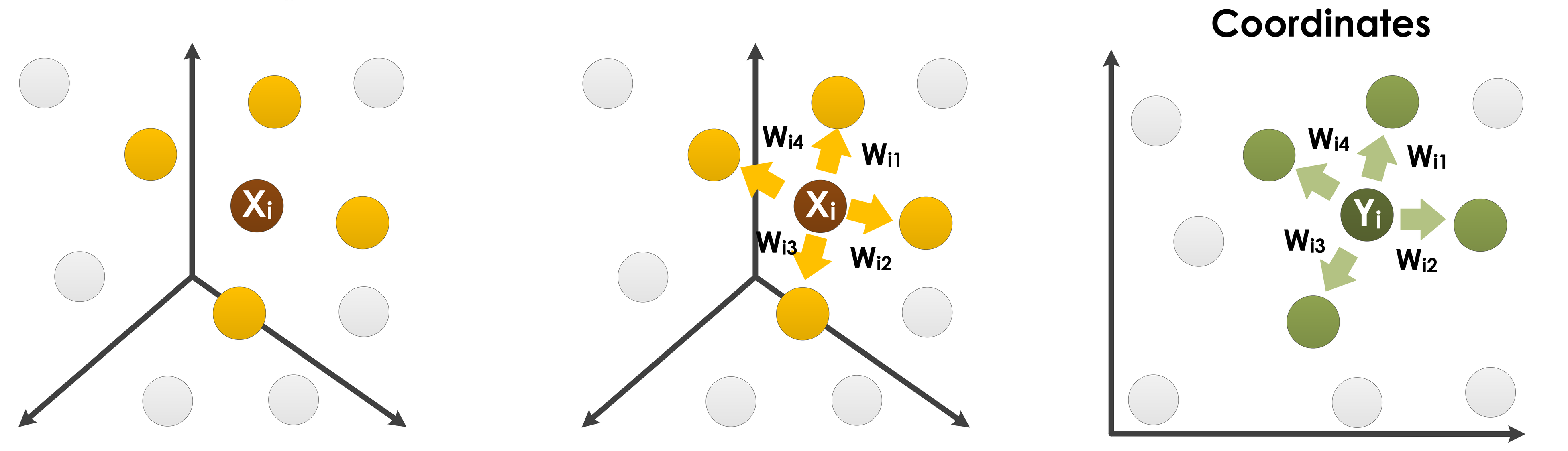
- Without utilizing the information of noisy speech
- Many to one or one to many issue
  - distorted speech frames from different SNRs are mapped to the same clean speech frame
  - different clean speech frames are mapped to the similar distorted speech frames

Proposed Method

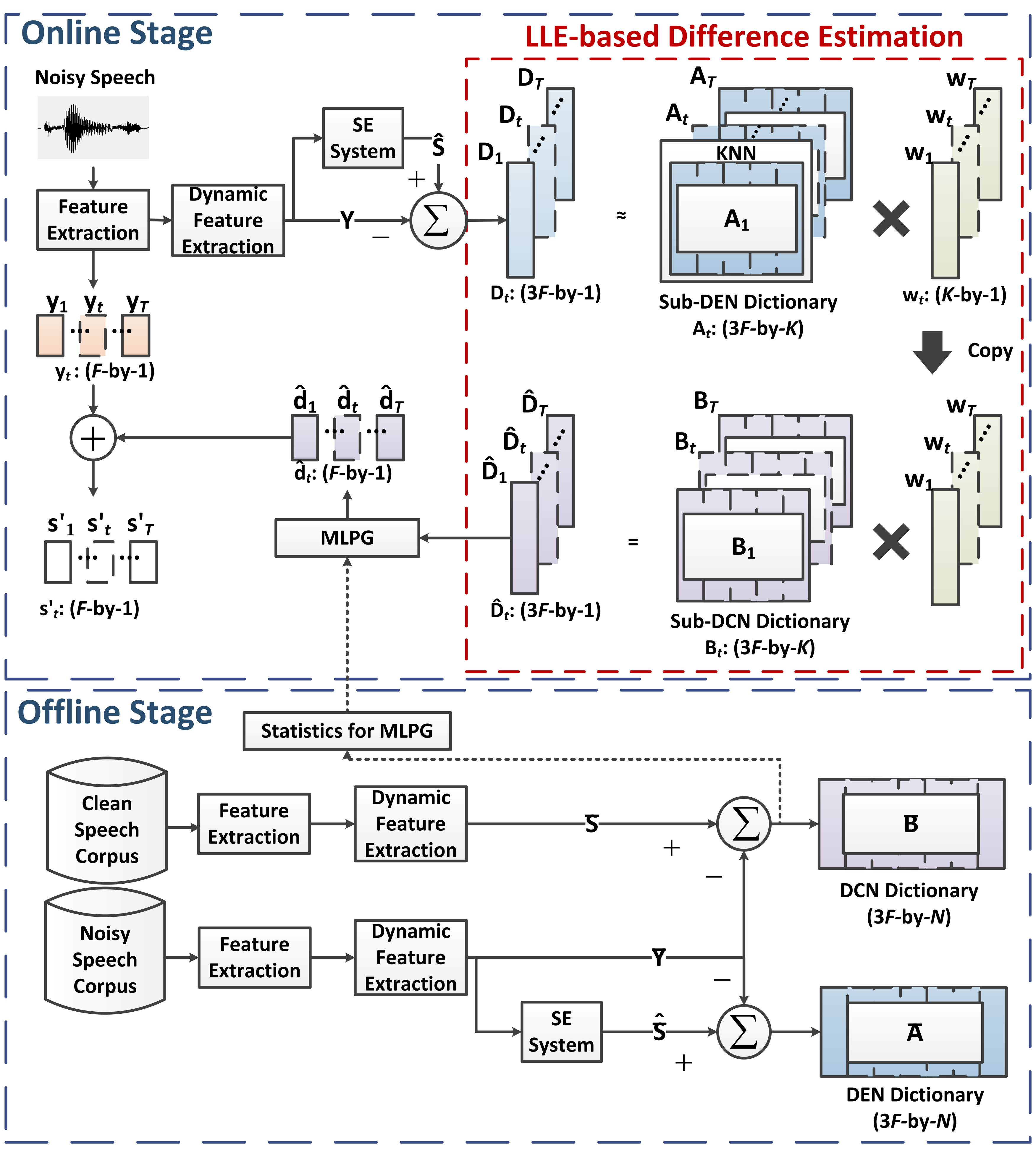
- Conversion of Wiener-like difference instead of speech
  - converting the difference of {SE-processed speech; noisy speech} to the difference of {clean speech; noisy speech}

System Framework

LLE Algorithm



LLE Difference Compensation



Experiments

- Corpus: Mandarin hearing in noise test (MHINT)
  - 300 utterances of a single speaker
  - 250 for training and 50 for testing
- Deep de-noising auto encoder (DDAE) system:
  - Structure: 7 hidden layers with 1200, 300, 300, 514, 300, 300, 1200 hidden nodes
  - Training data: 250 utterances mixed with 10~20 dB (5dB interval) car/two-talker noises
- Proposed (LDC) and directly conversion (DL) system:
  - Five-fold cross validation with 50 test utterances
  - Dictionary: 40 enhanced/clean utterances with SNR -10, 0, 10 dB car/two-talker noises
  - Testing data: 10 utterances with SNR -10, -6, -2, 0, 2, 6, 10 dB car/two-talker noises

Objective PESQ Score

PESQ						
Noise	Two-talker			Car		
Method	DDAE	DL	LDC	DDAE	DL	LDC
SNR10	2.21	2.22	**2.74**	1.96	2.03	**3.10**
SNR6	2.05	2.11	**2.44**	1.93	1.99	**2.88**
SNR2	1.93	1.97	**2.22**	1.89	1.92	**2.59**
SNR0	1.83	1.86	**2.08**	1.85	1.86	**2.43**
SNR-2	1.75	1.78	**1.95**	1.81	1.82	**2.28**
SNR-6	1.61	1.59	**1.74**	1.75	1.71	**2.02**
SNR-10	1.47	1.42	**1.56**	1.67	1.60	**1.82**
Ave	1.83	1.85	**2.10**	1.84	1.85	**2.44**

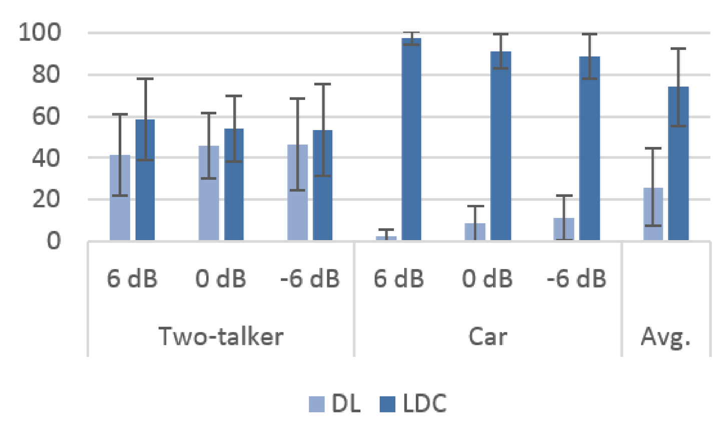
Objective SSNRI Score

SSNRI						
Noise	Two-talker			Car		
Method	DDAE	DL	LDC	DDAE	DL	LDC
SNR10	2.48	2.73	**3.99**	5.04	5.73	**7.59**
SNR6	5.76	6.08	**7.04**	8.17	8.91	**10.59**
SNR2	8.47	8.88	**9.51**	10.40	11.37	**12.63**
SNR0	9.66	10.12	**10.57**	11.40	12.34	**13.40**
SNR-2	10.46	11.03	**11.29**	12.00	13.05	**13.85**
SNR-6	11.38	**12.13**	11.39	12.34	13.74	**13.97**
SNR-10	11.51	**12.53**	11.12	12.22	**13.90**	13.45
Ave	8.53	9.07	**9.27**	10.23	11.29	**12.21**

Objective STOI Score

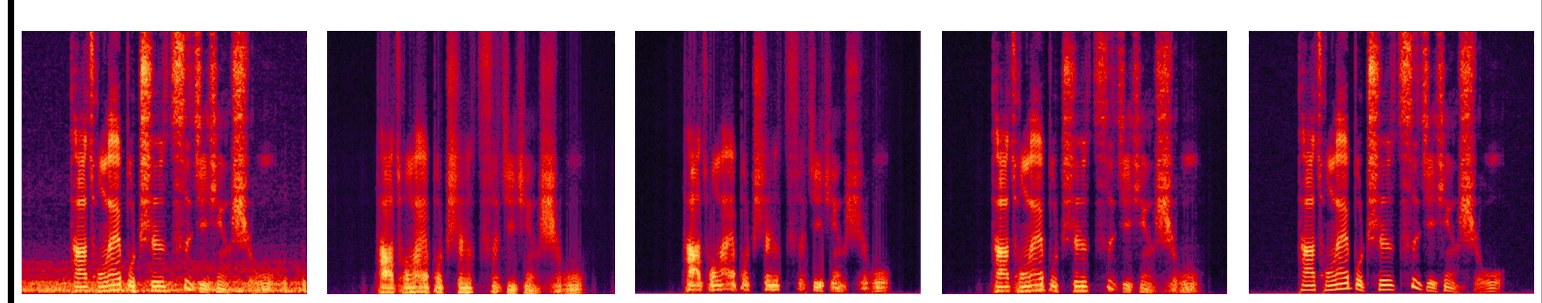
STOI						
Noise	Two-talker			Car		
Method	DDAE	DL	LDC	DDAE	DL	LDC
SNR10	0.88	0.83	**0.90**	0.85	0.80	**0.90**
SNR6	0.86	0.82	**0.88**	0.84	0.79	**0.88**
SNR2	0.84	0.80	**0.86**	0.83	0.78	**0.86**
SNR0	0.83	0.79	**0.84**	0.82	0.78	**0.85**
SNR-2	0.81	0.78	**0.82**	0.81	0.77	**0.83**
SNR-6	**0.78**	0.75	**0.78**	0.79	0.75	**0.80**
SNR-10	0.72	0.69	**0.73**	**0.76**	0.72	0.75
Ave	0.82	0.78	**0.83**	0.81	0.77	**0.84**

Subjective XAB Score



Spectrogram Results

NOISY DDAE DL LDC CLEAN



Conclusions

- Converting the difference between the SE-processed and noisy speech to the difference between the clean and noisy speech, and then compensating the noisy speech with the predicted difference, rather than directly converting the SE-processed speech to the clean speech
- Experimental results demonstrate that the proposed framework performs well in both objective and subjective evaluations