On the Role of Missing Data Imputation and NMF Feature Enhancement in Building Synthetic Voices Using Reverberant Speech

Dhananjaya Gowda[1], *Heikki Kallasjoki*[1], *Reima Karhila*[1], *Cristian Contan*[2], *Kalle Palomäki*[1], *Mircea Giurgiu*[2], *Mikko Kurimo*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland

[2] Department of Telecommunications, Technical University of Cluj-Napoca, Romania

[1]{dhananjaya.gowda; heikki.kallasjoki; reima.karhila; kalle.palomaki; mikko.kurimo}@aalto.fi

[2]{cristian.contan@bel; mircea.giurgiu@com}.utcluj.ro

I. Objective and Scope

- To build synthetic voices from reverberant speech data
- To study the usefulness of missing data imputation and NMF feature enhancement in improving the synthesized voices

II. Motivation

- Both missing data imputation and NMF feature enhancement show encouraging ASR results in the REVERB-2014 challenge [1]
- Reduction of baseline WERs by 43% and 34% on SimData and RealData, respectively, with clean training data.
- Reduction of baseline WERs by 18% and 17% on SimData and RealData, respectively, with multi-condition reverberant data for training.

III. Missing Data Imputation

Bounded Conditional Mean Imputation (BCMI) [2]:

- Model the distribution of clean speech $p(\mathrm{x})$ using a GMM
- Estimate the missing or unreliable part $\mathrm{x_u}$ of the noisy speech conditioning on the reliable part $\mathrm{x_r}$ as,

$$\hat{\mathrm{x}}_u = \int_{\mathrm{x}_u} \mathrm{x}_u p(\mathrm{x}_u|\mathrm{x}_r) d\mathrm{x}_u \quad (1)$$

- Use the upper bound: $\hat{\mathrm{x}}_u < \mathrm{x}_u^{obs}$

IV. NMF Feature Enhancement

Modeling reverberation in NMF:

$$\mathrm{Y} = \mathrm{RSA} \quad (2)$$

Y: $T_rC \times W$ stacked observation
R: $T_rC \times TC$ filter matrix
S: $TC \times N$ dictionary matrix
A: $N \times W$ activation matrix

- (RS)A: modeling with a reverberated dictionary
- R(SA): reverberating the NMF approximation

V. Feature Enhancement Algorithm

1. Estimate $\tilde{\mathrm{X}}$ using BCMI
2. Iteratively update A in $\tilde{\mathrm{X}} \approx \mathrm{RSA}$ with identity R
3. Filter A to suppress consecutive nonzero activations
4. Initialize R to contain filter $\frac{1}{T_f}\left[1 \ldots 1\right]$ on all channels
5. Iteratively update R in $\mathrm{Y} \approx \mathrm{RSA}$ with fixed A (under constraints $r_{t+1,b} < r_{t,b}, \quad \sum_{t,b} r_{t,b} = C$)
6. Iteratively update A in $\mathrm{Y} \approx \mathrm{RSA}$ with fixed R

- Then use $\hat{\mathrm{X}} = \mathrm{SA}$ and $\hat{\mathrm{Y}} = \mathrm{RSA}$ for feature enhancement, with a per-frame Wiener filter in the mel-spectral domain

VI. Experimental Setup

- GlottHMM based TTS system
- TTS features: Vocal tract LSF, source LSF, HNR, gain, F_0
- BCMI and NMF feature: Mel spectral energies derived from vocal tract LSFs
- Adaptation of average voice (TIMIT corpus) using reverberant and enhanced features
- Only vocal tract LSFs are enhanced/adapted.
- Other features are borrowed either from the average voice or from the reverberant data for synthesis

VII. Subjective Listening

- Synthesized samples: `http://research.ics.aalto.fi/speech/demos/reverb-synth-is2014/`
- Enhanced samples from the REVERB-2014 challenge: `http://research.spa.aalto.fi/speech/robust/kallasjoki-reverb14/`

VIII. Objective Evaluation

Mel-cepstral distortion (MCD):

$$d_{mcd} = \sqrt{2 \sum_{i=1}^{d} (c_i - \hat{c}_i)^2}. \quad (3)$$

where c and $\hat{c}$ are clean and reverberant/enhanced cepstra of dimension d.

Log-spectral distortion (LSD) (in dB):

$$d_{lsd} = \frac{1}{L} \sum_{k=1}^{L} 20 \times |\log_{10}(X[k]/\hat{X}[k])| \quad (4)$$

where $X[k]$ and $\hat{X}[k]$ are the clean and reverberant/enhanced mel magnitude spectra of L bands.

Frequency weighted segmental SNR (FWS) (in dB):

$$S_{fws} = 20 \sum_{k=0}^{L} W[k] \log_{10}\{X[k]/|X[k] - \hat{X}[k]|\}, \quad (5)$$

where $W[k] = X^{\gamma}[k] / \sum_{k=0}^{L} X^{\gamma}[k]$ is the frequency weight function. $X[k]$ and $\hat{X}[k]$ denote the mel magnitude spectra over L bands for clean and degraded/enhanced speech.

IX. Evaluation of Enhanced Samples

TABLE 1: *Objective scores comparing reverberant ('Reverb') and enhanced features ('BCMI' and final 'Enhanced') against the clean reference features. The lower the distortion measures (MCD and LSD) the better, and the higher the similarity score FWS, the better.*

	Meeting room			Lecture room		
Feature-type	MCD	LSD	FWS	MCD	LSD	FWS
Reverb	1.40	7.62	5.67	1.78	9.39	2.99
BCMI	1.47	8.06	5.67	1.75	9.11	3.65
Enhanced	**1.38**	**7.50**	**6.24**	**1.67**	**8.84**	**3.71**

X. Evaluation of Synthetic Samples

TABLE 2: *Objective scores comparing streams synthesized by different models against that of the clean speaker-dependent model.*

Cln-SpkDep vs diff. 'data-Model' combinations						
	Meeting room			Lecture room		
Data-Model	MCD	LSD	FWS	MCD	LSD	FWS
Rev-SpkDep	1.29	8.62	4.59	1.64	10.18	1.75
Rev-SpkAda	1.22	8.42	4.13	1.34	8.88	2.89
Enh-SpkAda	**1.18**	**7.36**	**5.66**	**1.27**	**7.94**	**4.18**

Cln-SpkDep vs Cln-AvgMod			
Data-Model	MCD	LSD	FWS
AvgMod-Cln	1.66	9.79	3.14

XI. Conclusions

- Adaptation of an average voice using reverberant data is better than building a speaker dependent voice from the reverberant data
- Adaptation of the average voice using enhanced features is better than direct adaptation using reverberant data
- Adaptation using the enhanced features reduces the distortion of synthesized samples when compared with clean speaker dependent voice samples

XII. Acknowledgements

- The Academy of Finland (Finnish Centre of Excellence in Computational Inference Research COIN, 251170)
- European Community's Seventh Framework Programme (FP7) under grant agreement no. 287678 (Simple4All).

References

[1] Heikki Kallasjoki, Jort F. Gemmeke, Kalle J. Palomäki, Amy V. Beeston, and Guy J. Brown, "Recognition of reverberant speech by missing data imputation and NMF feature enhancement," in *Proc. REVERB Challenge Workshop*, 2014.

[2] Ulpu Remes, "Bounded conditional mean imputation with an approximate posterior," in *Proc. INTERSPEECH*, 2013, pp. 3007–3011.

[3] R. Karhila, U. Remes, and M. Kurimo, "Noise in hmm-based speech synthesis adaptation: Analysis, evaluation methods and experiments," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 2, pp. 285–295, April 2014.