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
{dhananjaya.gowda; heikki.kallasjoki; reima.karhila; kalle.palomaki; mikko.kurimo}@aalto.fi
cristian.contan@bel.utcluj.ro; mircea.giurgiu@com.utcluj.ro

Abstract

In this paper, we study the role of a recently proposed feature enhancement technique in building HMM-based synthetic voices using reverberant speech data. The feature enhancement technique studied combines the advantages of missing data imputation and non-negative matrix factorization (NMF) based methods in cleaning up the reverberant features. Speaker adaptation of a clean average voice using noisy data is generally better than building a speaker dependent voice using the noisy data. In this paper, we show that the proposed feature enhancement technique can further improve the spectral match between the enhanced feature adapted voice and a clean speaker dependent voice.

Index Terms: dereverberation, missing data imputation, non-negative matrix factorization, GlottHMM, speech synthesis, speech enhancement

1. Introduction

Building synthetic voices using speech data recorded in realistic environments with natural degradations such as reverberation and background noise is a practical and challenging task. Enhancement of the degraded speech signal is an option, but is often accompanied with artifacts and sometimes even the loss of speaker characteristics which is not desired when we are trying to build personalized voices. Feature enhancement is a popular step in automatic speech recognition (ASR) for improving the system in degraded environments. The problem of degradation is less complex in ASR systems compared to synthesis systems, as the ASR systems rely only on features that represent the gross spectral envelope. Whereas synthesis systems need to model the excitation source along with the finer details of the spectrum so as to reconstruct the signal with least distortions. On the other hand, it has been demonstrated that HMM-based speech synthesis systems can provide some amount of robustness against degradations and varying recording conditions, when a high quality average voice is adapted using degraded data from a target speaker [1, 2]. The question this paper tries to address is whether the feature enhancement techniques can improve upon this robustness without introducing any artifacts of their own and still preserving the speaker characteristics.

Recently a new feature enhancement method combing missing data imputation and non-negative matrix factorization (NMF) techniques [3] was proposed for improving the performance of an ASR system in reverberant environment. The method tries to enhance the reverberant mel-spectral features first by estimating a mask for identifying unreliable regions, and then imputing the missing data using models built on clean data using bounded conditional mean imputation (BCMI) [4]. The missing data imputation step is followed by an NMF-based feature enhancement method extended with a filter model for reverberation that can be optimized in an unsupervised manner. It is shown that this new feature enhancement method improves significantly the performance of the ASR system under reverberant environment. In this paper, we study the usefulness of this features enhancement technique in building synthetic voices using reverberant speech data.

The paper is organized as follows: Section 2 gives an overview of the mask estimation, missing data imputation and the NMF-based feature enhancement techniques. Section 3 gives the details of the GlottHMM based speech synthesis used in the study for building synthetic voices. Section 4 describes the various experiments conducted along with their results and observations. A brief summary and conclusions are provided in Section 5.

2. Feature enhancement

The feature enhancement techniques used in this paper, proposed in [3], are briefly described in this section.

2.1. Missing data feature enhancement

For missing data feature enhancement, a *missing data mask* that classifies each spectro-temporal feature of the input signal as either reliable or corrupted by reverberation is required. We construct the mask by applying a method designed specifically for reverberant speech, originally proposed in [5]. The mask construction is based on the mel-spectral features, filtered with a band-pass modulation filter having 3 dB cutoff frequencies of 1.5 Hz and 8.2 Hz followed with an automatic gain control step. To form the mask, these smoothed features are thresholded using a channel-specific threshold value, automatically derived from a 'blurredness' heuristic that estimates the amount of reverberation present in the signal.

After the reliable regions have been identified, the corrupted regions are replaced by an estimate of the corresponding clean speech features, based on a bounded conditional mean imputation (BCMI) method proposed in [4]. A Gaussian mixture model is estimated to approximate the joint distribution $p(\mathbf{x})$ of clean speech features $\mathbf{x}$. Given the subsets $\mathbf{x}_{\mathrm{r}}$ and $\mathbf{x}_{\mathrm{u}}$ denoting the reliable and unreliable components of the features, respectively, the clean speech features can be estimated from the conditional distribution $p(\mathbf{x}_\mathrm{u} \mid \mathbf{x}_\mathrm{r})$ defined by the model. The BCMI approach further extends this estimate by using the corrupted observation $\mathbf{x}_\mathrm{u}$ as an upper bound for the enhanced features. In the method applied in this work, this upper bound is taken into account by deriving a parametric model of $p(\mathbf{x}_\mathrm{u} \mid \mathbf{x}_\mathrm{r})$ that approximates the distribution of the bounded features.

2.2. NMF-based feature enhancement

While the BCMI feature enhancement method described in Section 2.1 can be applied as a standalone system, we further complement it with a dereverberation approach based on non-negative matrix factorization (NMF), proposed in [3]. NMF is applied by assuming a compositional speech model of the form

$$\mathbf{Y} \approx \mathbf{RSA}, \tag{1}$$

where $\mathbf{Y}$ represents the reverberant observation, while $\mathbf{A}$ is a matrix of *activations* of non-negative clean speech basis atoms collected in the *dictionary* $\mathbf{S}$, and $\mathbf{R}$ represents the effect of reverberation. The columns of $\mathbf{S}$ and $\mathbf{Y}$ consist of short, overlapping spectrogram windows, with lengths of T and T_r frames for $\mathbf{S}$ and $\mathbf{Y}$, respectively, stacked to supervectors. The filter matrix $\mathbf{R}$ is a $T_rC \times TC$ matrix, where C is the number of mel channels, specifically constructed to perform a feature domain convolution within each such spectrogram with a filter of length T_f, where $T_r = T + T_f - 1$.

For feature enhancement, we use a fixed, predefined dictionary $\mathbf{S}$ constructed from clean speech data. The matrices $\mathbf{R}$ and $\mathbf{A}$ are obtained by minimizing, for each window t, the cost function

$$d(\mathbf{Y}_t, \mathbf{RSA}_t) + \lambda \left\|\mathbf{A}_t\right\|_1, \tag{2}$$

where the $d(\cdot,\cdot)$ term is the Kullback-Leibler divergence measuring the fit of the model to the observation, while the L^1-norm is intended to induce sparsity in the resulting factorization. An iterative algorithm consisting of multiplicative updates is used to minimize the cost function.

The BCMI feature enhancement method is used to produce an initial clean observation estimate $\tilde{\mathbf{X}}$. Initial activations $\mathbf{A}$ are derived from the factorization $\tilde{\mathbf{X}} \approx \mathbf{RSA}$ by keeping $\mathbf{R}$ fixed to the identity matrix, and updating $\mathbf{A}$ for I_1 iterations. The activation sequences of individual dictionary atoms, i.e., rows of $\mathbf{A}$, are filtered with a filter $H_A(z)$ to suppress successive activations characteristic of reverberant signals. The $\mathbf{R}$ matrix is then initialized to contain the constant filter $\frac{1}{T_f}[1 \ \ldots \ 1]$ for all channels, and updated based on $\mathbf{Y} \approx \mathbf{RSA}$ for I_2 iterations. As the multiplicative updates do not necessarily preserve the special structure of matrix $\mathbf{R}$, or result in a physically plausible filter, $\mathbf{R}$ is reinitialized after each iteration by averaging across all occurrences of each filter coefficient $r_{t,b}$, clamped to satisfy the condition $r_{t+1,b} < r_{t,b}$, and normalized so that $\sum_{t,b} r_{t,b} = C$. Finally, the $\mathbf{A}$ parameters are updated for further I_3 iterations, while keeping $\mathbf{R}$ fixed.

In order to allow for a reverberated clean speech dictionary atom activated in one window to "explain away" reverberant tails that occur early in successive windows it overlaps with, we replace the reconstruction $\mathbf{RSA}$ with the result of summing over its overlapping windows in the multiplicative update routines. This brings the model close to that of non-negative matrix factor deconvolution (NMFD) [6].

Using the final values of $\mathbf{R}$ and $\mathbf{A}$, we derive a per-frame Wiener filter based on the ratio of the clean speech reconstruction $\hat{\mathbf{X}} = \mathbf{SA}$ and the reverberant speech reconstruction $\hat{\mathbf{Y}} = \mathbf{RSA}$, after summing each across the overlapping windows. Final enhanced features are obtained by filtering the original observation with this filter.

3. Experimental setup

In this section, details of the GlottHMM speech synthesis system used for building speaker dependent and average voice models are provided. Details on the various parameters chosen for the NMF based feature enhancement are also provided.

3.1. GlottHMM based speech synthesis

GlottHMM based speech synthesis is a promising alternative to the popular HTS synthesis systems that use features extracted using a STRAIGHT vocoder [7–9]. It has been shown that GlottHMM based synthesis systems perform either comparable or better than the STRAIGHT-based systems in clean conditions. The source-filter modeling of speech by GlottHMM using inverse adaptive iterative filtering (IAIF) algorithm provides better modeling of the excitation source. It has also been demonstrated that it provides more flexibility for controlling different voice qualities such as soft, pressed, loud/lombard and creaky [10–12]. In view of this, we are using a GlottHMM system for our experiments in this paper. A comparison of performance with an HTS system using STRAIGHT features should make for an interesting study, but is not carried out in this paper.

The GlottHMM synthesis system used in this paper uses a 30 dimensional line spectral frequencies derived from a 30th order IAIF linear prediction (LP) analysis to model the spectral envelope or the vocal tract information. The excitation source is modeled using a 10th order LSFs representing the glottal flow spectral envelope, a gain parameter, a pitch or fundamental frequency (F_0) parameter which also embeds the voicing information, and five harmonic-to-noise ratio (HNR) parameters representing the aperiodicity information.

A quin-phone context-dependent hidden Markov model (HMM) based text-to-speech (TTS) system (also referred to as an HTS system in short) is built using the above parameters along with their delta and acceleration features with five states and one Gaussian mixture per state for each phone model. CSMAPLR method is used for speaker adaptation of the average voice [13].

3.2. BCMI-NMF based feature enhancement

The 30 dimensional LSF features are converted to a 61 dimensional mel-spectral features, by first converting LSFs to LPCs, LPCs to cepstral coefficients, a frequency warping to convert the cepstral coefficients to mel-cepstrals and then to mel-spectrum (linear and magnitude) using SPTK tools [14]. The enhanced mel-spectral features are converted back to LSFs before being fed to the speech synthesis system for adaptation. The other streams of the GlottHMM system are used as is from the reverberant speech for the adaptation step.

The 61 dimensional mel-spectral feature is used as input for the feature enhancement stage. Features for building the GMM models for BCMI and the NMF dictionary are also extracted in exactly the same procedure. The NMF dictionary consists of a total of 6520 randomly picked exemplars (2 per utterance). The window and filter length, sparsity coefficient and iteration count parameters of the NMF feature enhancement method of Section 2 are set to $T = 10$, $T_f = 20$, $C = 61$, $\lambda = 1$, $I_1 = 50$, $I_2 = 50$ and $I_3 = 100$, all based on small-scale experiments [3]. An FIR filter of the form $H_A(z) = 1-0.9z^{-1}-0.8z^{-1}-0.7z^{-1}$ is used to suppress the successive activation characteristics during the stage of learning the activation matrix A.

For the missing data mask estimation, parameter values of $\alpha = 19$, $\beta = 0.43$ and $\gamma = 1.4$ are chosen based on previous work [15]. The BCMI was performed using a 5-component GMM trained on a random 1000-utterance subset of the clean speech training set, with a time context of 3 consecutive frames for each window [4].

4. Experimental results

4.1. Data

The clean speech data used for experiments in this paper are that of a US male speaker 'bdl' part of the CMU-ARCTIC database [16]. The reverberant speech data is synthesized by convolving the clean speech signal separately with two different room impulse responses (RIR) from the AIR database [17]. The two different RIRs used are, one that of a low-to-moderately reverberant meeting-room with a reverberation time $T_{60} = 0.23s$, and the other of a heavily reverberant lecture-room with a $T_{60} = 0.78s$. The distance between the sound source and the microphone are 2.8m and 8.86m, respectively. The data for each of the clean, meeting-room and lecture-room scenarios contain a total of 800 phonetically balanced utterances that are used for training speaker-dependent voice models. All the 800 utterances for each of the three scenarios are used separately in generating the speaker-adapted voices by adapting a clean average male voice model. The average male voice model is built using data from all male speakers in the training subset of the TIMIT corpus [18]. This subset consists a total of 3260 utterances, 10 utterances each from 326 different male speakers. Also, all the 800 utterances after the feature enhancement stage are used for building the enhanced speaker-adapted voices.

4.2. Objective scores

Three different objective scores are used to evaluate the performance of the feature enhancement and speech synthesis systems.

Mel-cepstral distortion (MCD):

The mel-cepstral distortion between two cepstral vectors $\boldsymbol{c}$ and $\hat{\boldsymbol{c}}$ of dimension d is given by

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

Log-spectral distortion (LSD):

The log-spectral distortion (in dB) between two mel magnitude spectra $X[k]$ and $\hat{X}[k]$ of L bands is given by

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

Frequency weighted segmental SNR (FWS):

The frequency weighted segmental signal to noise ratio (SNR) (in dB) between two magnitude spectra is computed as

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

where $W[k] = X^{\gamma}[k]/\sum_{k=0}^{L} X^{\gamma}[k]$ is the frequency weight function, and $X[k]$ and $\hat{X}[k]$ denote the mel magnitude spectra over L bands for clean and degraded speech, respectively. A value of $\gamma = 0.2$ is used as proposed in [19]. It has been shown that this measure correlates well with the perceptual evaluation of speech quality (PESQ) measure, the industry standard for objective evaluation of speech quality [20, 21]. FWS is commonly used instead of PESQ because of highly priced licensing fees related to PESQ.

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**

4.3. Enhancement results

The performance of the BCMI step and the NMF based feature enhancement is studied using the objective measures listed above. The average MCD, LSD and FWS scores for reverberant, BCMI and the final NMF enhanced features as compared to the clean reference features is given in Table 1. A subset of 100 randomly selected utterances are used for computing the objective scores. It can be seen that for low-to-moderate reverberation as in the case of meeting room data, the intermediate BCMI enhanced features do not seem to provide much improvement over the reverberant data, while there is some improvement in the case of heavily degraded lecture room data. Nevertheless the final NMF enhanced features shows improvement for both the conditions.

4.4. Synthesis results

Results from a comparison of different voice models in terms of their spectral match with the clean speaker dependent voice model using different objective measures are given in Table 2. The different voice models studied are a speaker dependent voice built using reverberant data without any enhancement(Rev-SpkDep), a clean average voice model adapted using features extracted from reverberant data without and with the enhancement of the features (Rev-SpkAda and Enh-SpkAda, respectively). It can be seen that enhancement of the features significantly improves the performance in the case of lecture room data, while it shows a moderate improvement in the case of meeting room data. It is interesting to note how the average voice model matches against the clean speaker dependent voice, as given in the last row of Table 2. This is a good baseline score for interpreting other numbers in the table, as it gives a match between the speech or phonetic information between two clean systems. Any improvement on these numbers should mean an increased match in terms of speaker characteristics. The speaker dependent (Rev-SpkDep) and adapted (Rev-SpkAda) models built using heavily reverberant data (lecture room) are worse than the phonetic spectral match the average voice provides. Nevertheless speaker adaptation using enhanced features (Enh-SpkAda) bring the adapted voice much closer to the clean speaker dependent voice.

Table 2: Objective scores comparing streams synthesized by different models against that of the clean speaker-dependent model. 'Cln', 'Rev' and 'Enh' refer to clean, reverberant and enhanced, respectively. 'SpkDep' - speaker dependent, 'SpkAda' - speaker adapted, and 'AvgMod' - average model.

Cln-SpkDep vs diff. 'data-Model' combinations						
Data-Model	Meeting room			Lecture room		
	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

5. Conclusions

In this paper, we studied the role of missing data imputation and NMF-based feature enhancement techniques in improving the quality of a synthetic voice built using reverberant speech data. Feature enhancement techniques are popular in ASR for improving the performance of the system in noisy conditions. Speaker adaptation of an average voice model using noisy reverberant data improves the spectral match between the adapted voice and a speaker dependent voice built for the same speaker using clean data. It was shown in this paper that adaptation of the average voice model using features enhanced using the BCMI and NMF-based techniques can further improve the spectral match between the adapted voice and the clean speaker dependent voice. The usefulness of the feature enhancement method was demonstrated with the perceptually motivated frequency weighted segmental SNR measure that has been shown to correlate well with the PESQ standard measure. Also, traditional distortion measures such as mel-cepstral distortion and log-spectral distortion were used to demonstrate the improvement.

The feature enhancement techniques studied in this paper primarily focus on the spectral envelope. Other components of a synthesis model namely pitch, gain and source features also need to be enhanced for improving the voice quality further. A comparison of the NMF-based feature enhancement technique with other popular dereverberation techniques is our future interest. This paper presents a preliminary investigation on the suitability of the newly proposed NMF-based feature enhancement technique for the development of better synthesis models using reverberant speech data. The synthesized speech samples are available for listening at 'http://research.ics.aalto.fi/speech/demos/reverb-synthis2014/'. Informal listening tests on the various synthesized waveforms indeed show some improvement in speech quality, while still retaining the speaker similarity. A more formal listening test still needs to be carried out.

6. Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7) under grant agreement no. 287678 (Simple4All), and grants from the Academy of Finland (grant no. 136209, and Finnish Center of Excellence in Computational Inference Research COIN - grant no. 251170).

7. References

[1] J. Yamagishi, Z. Ling, and S. King, "Robustness of hmm-based speech synthesis," in *in Proc. Interspeech 2008*, 2008, pp. 581–584.

[2] 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.

[3] H. Kallasjoki, J. F. Gemmeke, K. J. Palomäki, A. V. Beeston, and G. J. Brown, "Recognition of reverberant speech by missing data imputation and NMF feature enhancement," in *Proc. REVERB Challenge Workshop*, 2014.

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

[5] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Communication*, vol. 43, no. 1-2, pp. 123–142, 2004.

[6] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, ser. Lecture Notes in Computer Science, C. G. Puntonet and A. Prieto, Eds. Springer Berlin Heidelberg, 2004, vol. 3195, pp. 494–499.

[7] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "Hmm-based speech synthesis utilizing glottal inverse filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 153–165, Jan 2011.

[8] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, May 2013.

[9] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," in *Proc. Int. Conf. Acoustics Speech and Signal Processing*, Las Vegas, USA, March 2008, pp. 3933–3936. [Online]. Available: http://www.wakayama-u.ac.jp/~kawahara/STRAIGHTtrial/

[10] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and Lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648–664, March 2014.

[11] T. Raitio, J. Kane, T. Drugman, and C. Gobl, "HMM-based synthesis of creaky voice," in *Interspeech*, Lyon, France, August 2013, pp. 2316–2320.

[12] J. Lorenzo-Trueba, R. Barra-Chicote, T. Raitio, N. Obin, P. Alku, J. Yamagishi, and J. M. Montero, "Towards glottal source controllability in expressive speech synthesis," in *Interspeech*, Portland, Oregon, September 2012.

[13] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for hmm-based speech synthesis and a constrained smaplr adaptation algorithm," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 66–83, Jan 2009.

[14] "Speech signal processing toolkit (sptk) version 3.7," December 2013. [Online]. Available: http://sp-tk.sourceforge.net/

[15] K. J. Palomäki, G. J. Brown, and J. P. Barker, "Recognition of reverberant speech using full cepstral features and spectral missing data," in *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP-06)*, 2006.

[16] J. Kominek and A. Black, "The CMU ARCTIC speech databases," in *5th ISCA Speech Synthesis Workshop*, Pittsburgh, PA, 2004, pp. 223–224. [Online]. Available: http://festvox.org/cmu_arctic/index.html

[17] M. Jeub, M. Schafer, and P. Vary, “A binaural room impulse response database for the evaluation of dereverberation algorithms,” in *Digital Signal Processing, 2009 16th International Conference on*, July 2009, pp. 1–5.

[18] J. S. Garofolo, et al., *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, Linguistic Data Consortium, Philadelphia, USA, 1993.

[19] Y. Hu and P. Loizou, “Evaluation of objective quality measures for speech enhancement,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[20] J. Tribolet, P. Noll, B. McDermott, and R. Crochiere, “A study of complexity and quality of speech waveform coders,” in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’78.*, vol. 3, Apr 1978, pp. 586–590.

[21] “ITU-T, Recommendation P.862 (02/2001) Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.”