The effect of speech enhancement in voice adaptation when building synthetic voices

Francis Del Prado¹, Yusuke Hioka^{1,*} and Catherine Watson^{2,†}

¹Department of Mechanical Engineering, University of Auckland, Auckland, 1142 New Zealand

²Department of Electrical and Computer Engineering, University of Auckland, Auckland, 1142 New Zealand

(Received 1 June 2017, Accepted for publication 1 August 2017)

Keywords: Speech synthesis, Voice adaptation, Speech enhancement, Beamforming, Wiener post-filter, Input SNR

PACS number: 43.72.Ja, 43.60.Fg [doi:10.1250/ast.39.150]

1. Introduction

Personalised synthetic voices have many applications for both computer agents and robots (e.g. [1]). Developing synthetic voices from scratch needs to be carefully handled as it requires a large amount of high quality speech samples and a specialist to record the speech material. This is a time consuming and expensive procedure. A solution is to transform an existing synthetic voice to a target voice (e.g. [2,3]). This technique creates the new voice by “seeding” an average voice model with a limited amount of speech from the target voice.

The voice adaptation system described here is to be used by a layperson in a domestic environment to create new voices for a healthcare robot [3]. The target users for the healthcare robot are the elderly, and the layperson would be someone familiar to them (e.g. family member). Studies have shown that the elderly do not perform well in intelligibility tests with synthetic speech compared to young adults (e.g. [4]), however speech intelligibility improves if the speech is from a familiar voice [5]. Therefore we plan to use voice adaptation techniques to create personalised voices for the healthcare robot.

The quality of the synthetic voice is dependent on the quality of the seeding speech material, so it is important the seeding material is free from background noise. To resolve the noise problem we used a speech enhancement system within the rapid voice adaptation process. The enhancement algorithm is based on the method proposed by Hioka *et al.* [6]. It consists of a combination of fixed beamforming before estimating power spectral density (PSD) of coherent signals in a room to compute a Wiener post-filter. For quality assurance a perception study was performed on the resulting synthetic speech, by 30 participants.

2. Rapid voice adaptation system

The rapid voice adaptation system (RVAS) [3] enables personalised voices on a healthcare robot, see Fig. 1. The text-to-speech system employed on the robot was MaryTTS [7]. MaryTTS provides tools for developing synthetic voices, including rapid voice adaptation, the voices are based on Hidden Markov models. The average voice for our RVAS system is based on 2 men and 2 women. The speech materials for the voice were acquired from the Mansfield Corpus [8].

The RVAS consists of two main components, a client on the robot, and a server. The person to be recorded for the personalised voice uses a purpose built application on the robot to record audio data. The application guides the user through recording the first 150 sentences of the Mansfield corpus [8]. The default microphone and speakers on the robot are automatically selected, but this can be modified. The RVAS application enables users to test the microphone levels, listen to their recordings, and re-record if necessary. Once all 150 sentences are completed, the speech data is submitted to the robot server via the application. The computational intensive process of creating a personalised voice with the newly acquired seeding material is performed on the server using the MaryTTS voice building tools. Once the new voice has been created, the healthcare robot can request the synthetic voices from the server and configure the newly adapted synthetic voice for use on the robot. All speech used in the RVAS is from New Zealand English speakers, including the seeding material.

3. Speech enhancement

3.1. Algorithm

Figure 2 depicts the block diagram of the speech enhancement algorithm applied in this study [6]. The algorithm is based on *beamforming with post-filter*, a well-known framework for its effectiveness in practical applications of speech enhancement. The algorithm consists of two steps, beamforming (③ in Fig. 2) and Wiener post-filter (④ in Fig. 2), both of which separate signals arriving from different directions.

Beamforming [6] plays a role as a spatial filter and realises the spatial selectivity of microphone arrays. The output of a beamformer is given by

$$Y(\omega, \tau) = \sum_{m=1}^M W_m(\omega) X_m(\omega, \tau), \quad (1)$$

where $W_m(\omega)$ is the weight of the beamformer for each microphone m , and $X_m(\omega, \tau)$ is the short-time Fourier transform of the m -th microphone observation. ω and τ denote the frequency and frame indices, respectively. In this study, delay-and-sum beamforming [6] is utilised to determine the beamformer's weights.

In order to boost the effect of speech enhancement, a post-filter is applied to the output of the beamforming. The most commonly used design of the post-filter is the Wiener filter defined by

*e-mail: yusuke.hioka@ieec.org

†e-mail: c.watson@auckland.ac.nz

Fig. 1 A user operating the RVAS client interface on the healthcare robot.

Fig. 2 The speech enhancement algorithm.

$$H(\omega, \tau) = \frac{\phi_S(\omega, \tau)}{\phi_S(\omega, \tau) + \phi_N(\omega, \tau)}, \quad (2)$$

where $\phi_S(\omega, \tau)$ and $\phi_N(\omega, \tau)$ are the power spectral densities (PSDs) of the target signal and interference, respectively, which are estimated by *PSD estimation in beamspace* [9]. By applying $H(\omega, \tau)$ to the output of the beamformer $Y(\omega, \tau)$, the target speech in the output of the beamformer is emphasised.

3.2. Modelling the room acoustics and noise of a domestic environment

The proposed speech enhancement module for RVAS is still a proof of concept. Eventually the user will be recording speech using the healthcare robot equipped with a microphone array. Since the performance of a microphone array is often affected by the acoustical property of the environment, such acoustical property of a typical domestic environment should be taken into account in our prototype investigation. To that end, in this preliminary analysis, the microphone array observation (② in Fig. 2) was generated by convolving a sound source (either clean speech or interfering noise) and the impulse response between the source position (① in Fig. 2) and the microphone array. The impulse response was measured in the listening room of the Acoustics Laboratory, University of Auckland. The room was designed to replicate a typical living room. The dimension and reverberation time (T_{60}) of the room are approximately $5.5 \times 8.3 \times 2.4$ m and 0.5 s at 1 kHz, respectively. The microphone array used had 8 linearly spaced (4 cm inter-spacing) microphones. Both the microphone array and loudspeaker were located in the centre of the room at 1.0 m height. Once the observation of both the target speech and interfering noise were generated, they were added together to provide a simulated microphone array observation. This approach offered a flexible and time saving method to achieve a realistic recording for voice training without having to re-record the corpus.

3.3. Applying speech enhancement to rapid voice adaptation

To investigate building voices using speech enhancement, the speech enhancement module was applied to the microphone observation (as per Fig. 2), which was generated using either pink noise or babble noise as the interfering noise. The resulting speech files were then uploaded to the robot speech server for rapid voice adaptation and a new adapted synthetic voice was created and ready for use in MaryTTS on the robot.

Initially, the quality and intelligibility of the adapted synthetic voices were poor. The additive noise was increasing the amplitude of the sound files, causing clipping in resulting signals. This caused problems with the pitch calculation. To alleviate the problem, the speech data was normalised to 70% of its maximum magnitude after applying speech enhancement but before being uploaded to the server for voice adaptation. This resulted in an acceptable adapted voice, with reduced noise.

4. Synthetic speech perception study

We carried perception tests to assess the efficacy of the proposed speech enhancement method in reducing noise in the rapid voice adaptation process. The test was delivered via an online survey to increase the potential pool of participants. Thirty participants did the test. Of these 25 were aged between 18 and 30 years, and the remaining 5 were aged between 30 and 60 years. Twenty-four participants self reported they had no hearing loss, the four were unsure. Twelve participants used loud-speakers to do the perception test, the remaining 18 used headphones, and 26 participants said they were familiar with synthetic speech.

The perception study was organised into two sections: one to assess speech discrimination and the other to assess speech quality.

4.1. Speech discrimination

The voices tested for this section were based on two factors: the degree of speech enhancement before voice adaptation and the additive noise contaminant before noise removal. The various voices investigated came from the seeding material of a single male speaker. The noise sources selected were pink noise and babble speech at an input SNR of 10 dB. Four voices were investigated.

Speech discrimination was investigated using an auditory ABX test to establish whether participants could hear any difference between adapted synthetic voices. The participants listened to three audio clips, A, B and X, where A and B are two different clips and X is either A or B. The participants then identified whether X is A or B. If they are correct they can successfully discriminate between A and B. If they are not there is no perceptible difference between A and B. The survey was designed so that the participant could only listen to each audio sound file once and allowed a third option if the participant could not decide whether X was A or B in a particular question. All audio clips were synthetic speech generated from MaryTTS. All conditions were assessed with three different sentences (11–14 words in length). The sentences were: “Each guest who came into the breakfast room was bombarded with the wonderful news,” “The second bells rung she urged, knowing perfectly well she was telling

her story,” and “Hello my name is Charlie and I am a healthcare robot.”

There were three different ABX combination types that the participants were asked to discriminate: (1) no overlapping conditions e.g. a voice created from speech with babble noise condition enhanced with beamforming and the Wiener post-filter may be compared to a voice created from pink noise then enhanced with beamforming, (2) different noise condition (e.g. up to ④ in Fig. 2), but with pink noise vs. babble noise, (3) different enhancement method (i.e. beamformer (③ in Fig. 2) vs beamformer and Wiener post-filter (④ in Fig. 2).

4.1.1. Results

When there were no similarities in preparation conditions of the voices pairs, 28/30 participants could reliably distinguish between the voices, the remaining two were incorrect. When looking at the impact of noise type, only 10 participants could correctly discriminate between the A and B samples, 6 were incorrect, and 14 could not decide. In contrast, when looking at the impact of speech enhancement methods, 27/30 participants could distinguish between the adapted voices, only three were incorrect. This suggests that whilst the type of noise contaminating the seeding speech does not create a discernible difference in the adapted voices, the speech enhancement process does. Thus, appending the Wiener post-filter to the beamformer output adds a perceptible difference to the signal. However, from these results alone it is impossible to determine if the post-filter made a positive or negative impact on the signal.

4.2. Speech quality

The second part of the test assessed the quality of the adapted speech created from seeding material which used speech enhancement. The participants evaluated on five point scales, the quality (poor to excellent), pleasantness (unpleasant to very pleasant), and listening effort (exhausting to very easy) of the voices. There were eleven adapted voices in total that were investigated for this section. The factors investigated were the two types of additive noise; pink and babble noise (PN and BB respectively), which were added at different levels of input SNR (10 dB and 0 dB). We also investigated the impact of speech enhancement to the overall perception, so no speech enhancement, fixed beamforming (up to ③ in Fig. 2, and henceforth referred to as BF) and the combination of fixed beamforming and Wiener post-filter (up to ④ in Fig. 2, and henceforth referred to as WF), were evaluated. Eleven voices were investigated as summarised in Table 1.

4.2.1. Results

The median speech quality results are tabulated in Table 1. The medians for the pleasantness and listening effort rating were exactly the same as quality, except in two instances. These were for beamforming and 10 dB pink noise, where pleasantness and listening effort had a median of 3, and for beamforming and 0 dB babble, where listening effort had a median of 1.5.

Statistical analysis was performed on the data using the Wilcoxon signed rank test based on the three factors: degree of speech enhancement, impact of noise level (input SNR) and the impact of noise type. The significance level was set to

Table 1 The median speech quality rating for different experimental conditions.

Adapted speech	Median quality rating	
	SNR: 10 dB	SNR: 0 dB
Pink noise (PN)	3	1
Beamforming & pink noise (BF+PN)	3.5	2
Wiener filter & pink noise (WF+PN)	2	2
Beamforming & babble speech (BF+BB)	3	2
Wiener filter & babble speech (WF+BB)	2	1
CLEAN	3

Fig. 3 Boxplot showing overall quality scores of voices with pink noise (a) input SNR: 10 dB, (b) input SNR: 0 dB.

$\alpha = 0.05$ and the hypothesis for the statistical tests were performed as two-tailed or non-directional signed rank tests. Only the speech quality results are presented, since there was little difference between these and the statistics of the pleasantness and listening effort results.

We investigated the degree of speech enhancement at different input SNR to determine whether appending a Wiener post-filter to the output of the beamformer had a positive or negative impact to the quality scores. We expect that a higher degree of speech enhancement will have the higher median scores of the adapted synthetic voice but we expect clean and noisy speech to be in extremities, i.e. $CLEAN > WF+PN > BF+PN > PN$.

Figure 3(a) displays the box plots of the speech quality scores of these voices at an 10 dB input SNR. Unexpectedly for the input SNR of 10 dB the Wiener post-filter and beamformer combination (WF+PN 10 dB (median = 2)), had the worst quality rating compared to the other voices (see Fig. 3(a)) and these differences were statistically significant: PN 10 dB (median = 3, $z = -3.19$) BF+PN 10 dB (median = 3.5, $z = -4.00$) and CLEAN (median = 3, $z = 3.82$). Figure 3(b) displays the box-plots of the quality scores at 0 dB input SNR. At 0 dB input SNR the pink noise without speech enhancement is very detrimental to the quality perceived, 28/30 participants gave it the lowest quality

rating. When speech enhancement was applied we can see that it makes an improvement to the quality ratings. The difference of the quality scores of PN 0dB voice (median = 1) compared to the other three voices is statistically significant: BF+PN 0dB (median = 2, $z = 4.0212$) and WF+PN 0dB (median = 2, $z = 3.6874$) and CLEAN (median = 3, $z = 4.5963$), but the rating is only going from “poor” to “below average.”

Next we investigated whether there were any significant trends associated with the noise level. We expect that a higher input SNR of the additive noise will result to higher quality scores for the adapted synthetic voice, and we expect the clean and noisy speech to be at the extremities, i.e. Clean speech > Additive noise + speech 10dB input SNR > Additive noise + speech 0dB input SNR. Thus WF+PN 10dB was compared to WF+PN 0dB, and BF+PN 10dB was compared to BF+PN 0dB (see Table 1 for abbreviation definitions). We expected the quality scores when the additive noise differed in SNR would be different, when the speech enhancement method was kept constant. However the difference was only significant between the voices BF+PN 10dB (median = 3.5) and BF+PN 0dB (median = 2) with a z -score = 4.2259. Regardless of the noise level, the quality scores for voices enhanced with both the beamformer and the Wiener post-filter remained poor.

In the final investigation we compared the noise conditions babble and pink noise. Due the participants inability to distinguish between voices contaminated pink noise or babble in the discrimination test (see Sect. 4.1.1.) we expected the voices to have a similar quality measure scores, i.e. speech + pink noise = speech + babble speech. The four comparisons investigated were: WF+PN 10dB vs. WF+BB 10dB, BF+PN 10dB vs. BF+BB 10dB, WF+PN 0dB vs. WF+BB 0dB, BF+PN 0dB vs. BF+BB 0dB. At input SNR 10dB there were no significant differences, as expected. However at input SNR 0dB there were significant differences in the quality measures between the voices WF+PN 0dB (median = 2) and WF+BB 0dB (median = 1, returning no z score) and between the voices BF+PN 0dB (median = 2) and voice BF+BB 0dB (median = 2, returning no z score).

5. Conclusion

In this study, we have implemented a speech enhancement algorithm and merged it with an existing voice adaptation system developing a testing platform which will ultimately create personalised adapted synthetic voices on a healthcare robot. The rapid voice adaptation system is susceptible to noise, and we investigated the use of a speech enhancement algorithm to reduce the noise in the recordings of the seed material for rapid voice adaptation. The voices developed used a conditioned training corpus to recreate recordings observed in a domestic environment. A perception

test with 30 participants was used to evaluate the speech enhanced synthetic voices. We found that the Wiener post-filter had a negative impact on the quality of the adapted voice for an input SNR of 10dB, however if the speech enhancement only used the beamformer the speech quality was at least as good as clean speech. When the input SNR was 0dB, both methods of speech enhancement improved the speech quality score when the additive noise was pink, and using the beamformer for babble noise. But in all three cases the improvement was only from “poor” to “below average.” It would be useful to get a better understanding on the behaviour of the process at the SNR levels between 0dB and 10dB. It should also be addressed that there would be a trade-off between the amount of noise reduced and that of distortion in the desired speech caused by the process of speech enhancement. So whilst more noise is being removed from speech, crucial aspects of the speech signal are also being modified which results in a poor quality adapted voice. The reasons behind this need to be further investigated.

References

- [1] C. Jayawardena, I. H. Kuo, U. Unger, A. Igic, R. Wong, C. I. Watson, R. Stafford, E. Broadbent, P. Tiwari, J. Warren, J. Sohn and B. A. MacDonald, “Deployment of a service robot to help older people,” *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5990–5995 (2010).
- [2] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi and K. Oura, “Speech synthesis based on hidden Markov models,” *Proc. IEEE*, **101**, 1234–1252 (2013).
- [3] S. Jain, “Towards the creation of customised synthetic voices using hidden markov models on a healthcare robot,” *Masters Thesis, University of Auckland* (2015).
- [4] R. W. Roring, F. G. Hines and N. Charness, “Age differences in identifying words in synthetic speech,” *Hum. Factors: J. Hum. Factors Ergon. Soc.*, **49**, 25–31 (2007).
- [5] P. Souza, N. Gehani, R. Wright and D. McCloy, “The advantage of knowing the talker,” *J. Am. Acad. Audiol.*, **24**, 689–700 (2013).
- [6] Y. Hioka and K. Niwa, “Estimating power spectral density for spatial audio signal separation: An effective approach for practical applications,” *Acoust. Sci. & Tech.*, **38**, 175–184 (2017).
- [7] M. Schröder and J. Trouvain, “The German text-to-speech synthesis system mary: A tool for research, development and teaching,” *Int. J. Speech Technol.*, **6**, 365–377 (2003).
- [8] C. I. Watson and A. Marchi, “Resources created for building New Zealand English voices,” *Proc. 15th Australas. Int. Conf. Speech Science and Technology*, pp. 92–95 (2014).
- [9] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa and Y. Haneda, “Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain,” *IEEE Trans. Audio Speech Lang. Process.*, **21**, 1240–1250 (2013).