# HMM-BASED SPEECH SYNTHESIS ADAPTATION USING NOISY DATA: ANALYSIS AND EVALUATION METHODS

*Reima Karhila, Ulpu Remes and Mikko Kurimo*

Department of Information and Computer Science, Aalto University School of Science, Finland

reima.karhila@aalto.fi, ulpu.remes@aalto.fi, mikko.kurimo@aalto.fi

## ABSTRACT

This paper investigates the role of noise in speaker-adaptation of HMM-based text-to-speech (TTS) synthesis and presents a new evaluation procedure. Both a new listening test based on ITU-T recommendation 835 and a perceptually motivated objective measure, frequency-weighted segmental SNR, improve the evaluation of synthetic speech when noise is present. The evaluation of voices adapted with noisy data show that the noise plays a relatively small but noticeable role in the quality of synthetic speech: Naturalness and speaker similarity are not affected in a significant way by the noise, but listeners prefer the voices trained from cleaner data. Noise removal, even when it degrades natural speech quality, improves the synthetic voice.

***Index Terms***— Speech Synthesis, Adaptation, Noise robustness, Evaluation, Feature extraction

## 1. INTRODUCTION

A text-to-speech (TTS) system based on speaker-adaptive hidden Markov model (HMM) synthesis [1] allows building synthetic voices that mimic any person with little effort. A good quality average voice model trained from a large population of speakers can be adapted to a new speaker with just a few minutes of recorded speech [2, 3].

This paper deals with noise in speech data that is used to personalise a TTS voice. In mobile voice manipulation applications and in found data cases (i.e. using archived material for new purposes) it is necessary to use data recorded in conditions that are far from studio level quality and might include intrusive background noise. Background noise is assumed to be a problem for two reasons: the noise can mask speaker characteristics that synthesis adaptation should learn, and adaptation can learn to reproduce the noise in the synthesised voice. Speech enhancement can remove noise to a certain extent, but aggressive noise suppression also distorts the speech signal.

Speaker-adaptive HMM-based speech synthesis has demonstrated robustness to quality variations in recording conditions when the adaptation data is collected from various sources, such as in [4]. The data selection process in [4] included removing recordings with background noise such as music or applause. This is a typical approach in speech synthesis systems, and works well when enough clean data is available.

In this work, we investigate learning speaker-adaptive transformations for HMM-based speech synthesis from noisy data, and discover that noise in training data plays a relatively small but noticeable role in the overall quality of the resulting synthetic speech. We show that by automatically removing noise from the training data, we can obtain high-quality adapted voices even when the lack of high-quality data prevents the traditional data selection approach. Our goal is also to find evaluation methods that allow reliable measurement of synthesis quality and listener preference in adverse conditions, and we develop a listening test focused on noisy data as well as implement perceptually motivated objective measures.

The remainder of this paper is organised as follows: In Section 2 we describe our synthesis system and speech and noise databases. We also introduce a new formulation of ITU-T recommendation 835 [5] as a subjective test for speech synthesis. We also propose to use frequency-weighted signal-to-noise ratio (fwS) as an objective measure for speech enhancement and synthesis evaluation. In Section 3 we investigate the effects of noise in various stages of adaptation: preprocessing with exemplar-based speech enhancement proposed in [6], feature extraction with STRAIGHT, and synthesis using linear regression transforms for speaker adaptation. Then in Section 4 we present a listening test developed for this work and its results, and in Section 5 we present our conclusions.

## 2. SETUP

### 2.1. Model training and adaptation

Male and female average voice models were trained from the Finnish PERSO synthesis database using the methods and tools of the EMIME 2010 Blizzard Entry [7] based on the HTS speech synthesis toolkit. In short, context-dependent multi-space distribution hidden semi-Markov models (MSD-HSMM) were trained on acoustic feature vectors comprising of STRAIGHT-analysed mel-generalised cepstral coefficients (MCEP), fundamental frequency, and aperiodicity features, computed from 16kHz speech signals. Speaker-adaptive training was applied to create speaker-adaptive average voice models.

The speaker adaptation experiments were carried out with data from the EMIME TTS corpus [8] using 105 sentences both from three male and three female native Finnish speakers. Noisy versions of the data were prepared by adding noise from NOISEX-92 [9] to the EMIME sentences at designated signal-to-noise ratios (SNR). For each utterance, the average energy was calculated for both clean speech and a noise signal of the same length, and the noise was scaled to match the desired SNR. Gender-dependent average voice models were adapted using constrained structural maximum a posteriori linear regression (CSMAPLR) [10]. CSMAPLR applies structural MAP criterion to learning constrained maximum likelihood linear regression (CMLLR) [11] transformations for the acoustic model mean and variance parameters.

### 2.2. Subjective evaluation methods

The standard MOS tests, where listeners listen to natural and synthesised speech samples and judged them based on their naturalness and speaker similarity on a subjective scale from one to five, were found to be misleading when there is background noise in the test samples. Preliminary tests revealed that listeners have widely varying approaches to noise, and it was not possible to find statistical differences even between strongly differing samples. In particular, background noise seems to mask synthesis artifacts and give more favourable judgements to the noisier samples.

The subjective evaluation method proposed in this work is based on MOS-tests of naturalness and speaker similarity, and the ITU-T recommendation 835 for testing telecommunication systems with noise suppression algorithms [5]. In this test method, listeners listen to the same sample three times, each time answering a different question. The variations geared towards speech synthesis evaluation are show in Table 1.

The ITU-T recommendation specifies using 4s chunks of speech, whereas the utterances used here were generally shorter. Thus, in our approach whole sentences are played with beginning and end silences cut off.

### 2.3. Objective evaluation methods

Traditional methods for speech synthesis evaluation are concentrated on the quality of output speech. In the case of noisy training data, some background noise will be present in some speech samples. A typical objective evaluation measure used in speech manipulation is the mel-cepstral distortion (MCD), which is also used in this work. However, as it is not well known, how MCD responds to a combination of noise and synthesis, we use it in conjunction with a more perceptually motivated evaluation method. **Frequency-weighted segmental SNR (fwS)** [12] is an improved version of segmental SNR [13]. It was found to correlate well with the industry standard objective evaluation method PESQ [14], with substantially smaller implementation and computational cost [12].

After frame-based normalisation is applied to FFT spectra of both the test and reference signal, mel filter-bank values $X$ and $\hat{X}$ are calculated with VOICEBOX [15]. The fwS measure is calculated from these as:

$$fwS = \frac{10}{M} \times \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m) log_{10} \frac{|X(j,m)|^2}{(|X(j,m)|-|\hat{X}(j,m)|)^2}}{\sum_{j=1}^{K} W(j,m)} \quad (1)$$

where $|\hat{X}(j,m)|$ is the test signal value in the $j$th mel filter channel at the $m$th frame, $|X(j,m)|$ is the reference signal value in the same mel channel, $W(j,m) = |X(j,m)|^{\gamma}$ with $\gamma = 0.2$ as proposed in [12], and $K$ is the number of filter banks, $M$ is the total number of frames in the signal.

The SNR in each frame is bound to [0,35] dB range. SNRs above 35dB are not perceptually much different. Similar to the ITU-T recommendation [5], in this work the fwS value is calculated without the beginning and end silences, using only 2s chunks taken from the middle of each utterance.

## 3. ANALYSIS OF NOISE IN HMM-TTS ADAPTATION

### 3.1. Speech enhancement

Given adaptation data where noise-corrupted samples exist, the standard approach used in speech synthesis is to discard utterances whose quality is considered too low [4, 2], but this is not an attractive option when the amount of high-quality adaptation data is small. Instead, speech enhancement methods developed for noise-robust automatic speech recognition (ASR) can be applied to calculate a clean speech estimate from the noise-corrupted observations.

Using speech enhancement in conjunction with synthesis adaptation has not been widely studied. There are numerous ways of removing noise from corrupted speech signals. Average voice training in earlier work [16] showed that single-channel non-negative matrix factorisation (NMF) based speech enhancement method proposed in [6] is efficient in improving the overall quality of the used average voice. However, [16] did not analyse the improvements in any methodological way. We will investigate this properly now with factory and babble noises.

The exemplar-based NMF code books for noth noise types were randomly sampled from the Finnish SpeeCon corpus and the babble and factory floor data. The clean speech and noise samples selected for adaptation data were not present in the code book training data. The parameters were optimised based on mel-cepstral distortion (MCD) between the original and separated speech samples of a subset of speakers. The enhanced data was then used in adaptation in a normal fashion.

The leftmost result columns in Table 2 show the measured MCD and fwS scores for both original (top) and enhanced natural speech (bottom). The effects of NMF-based speech enhancement are clear. Both objective measures show that for good quality signals (SNR 20 babble) the quality degrades, whereas for very noisy cases (SNR 5 factory or babble) the quality is significantly improved.

**Table 2**: Averaged fwS and MCEP distortion measures for 3 male and 3 female speakers, for original training data, vocoder-analysed and resynthesised training data and synthetic data generated by adapted synthesis models.

| | | Original training data | | Vocoder-resynthesised training data | | HTS-synthesised test data | |
|---|---|---|---|---|---|---|---|
| Noise | SNR | fwS | MCD | fwS | MCD | fwS | MCD |
| Clean | - | 35.00 | 0 | 14.77 | 1.14 | 9.60 | 1.87 |
| Babble | 20 | 19.65 | 1.32 | 13.05 | 1.92 | 8.93 | 2.10 |
| | 10 | 12.38 | 2.19 | 10.19 | 2.88 | 8.47 | 2.37 |
| | 5 | 9.13 | 2.70 | 8.23 | 3.35 | 8.14 | 2.58 |
| Factory | 10 | 9.54 | 2.88 | 8.28 | 3.56 | 7.80 | 2.69 |
| | 5 | 6.58 | 3.44 | 6.15 | 4.15 | 7.33 | 3.02 |
| Machine Gun | 0 | 20.66 | 1.06 | 12.71 | 1.69 | 8.93 | 2.02 |
| Enhanced Babble | 20 | 18.64 | 1.35 | 13.16 | 1.91 | 8.94 | 2.06 |
| | 10 | 12.65 | 1.95 | 10.54 | 2.63 | 8.49 | 2.24 |
| | 5 | 9.73 | 2.22 | 8.78 | 3.12 | 8.22 | 2.37 |
| Enhanced Factory | 10 | 10.55 | 2.19 | 9.09 | 3.02 | 7.84 | 2.44 |
| | 5 | 8.17 | 2.40 | 7.53 | 3.41 | 7.52 | 2.61 |

### 3.2. Feature extraction

Numerous feature extraction methods exist for improving the quality of ASR in noisy surroundings. Many of these techniques normalise the speech signal and are therefore difficult to apply to TTS systems,

**Table 1**: Questions used in the subjective evaluation task 1

| | |
|---|---|
| (a) Play the sample and attending ONLY to the SPEECH SIGNAL, select the category which best describes the sample you just heard.<br>the SPEECH SIGNAL in this signal was<br>**5.** Completely natural<br>**4.** Quite natural<br>**3.** Somewhat unnatural but acceptable<br>**2.** Quite unnatural<br>**1.** Completely unnatural | (c) Play both samples, and attending ONLY to the SPEECH SIGNAL, select the category which best describes the second sample to the reference sample.<br>The voices in the SPEECH SIGNALS of the samples sounded:<br>**5.** Exactly like the same person<br>**4.** Quite like the same person<br>**3.** Somewhat different but recognisable as the same person<br>**2.** Quite like a different person<br>**1.** Like a totally different person |
| (b) Play the sample and attending ONLY to the BACKGROUND, select the category which best describes the sample you just heard.<br>the BACKGROUND in this signal was<br>**5.** Clean<br>**4.** Quite clean<br>**3.** Somewhat noisy but not intrusive<br>**2.** Quite noisy and somewhat intrusive<br>**1.** Very noisy and very intrusive | (d) Play the reference sentence. Then play both sample sentences. Considering the OVERALL QUALITY of the signal, select the one you would prefer to represent the reference voice in applications like mobile devices, video games, audio books etc.<br>Regarding the OVERALL QUALITY<br>**A.** First sample is better<br>**B.** Second sample is better<br>**C.** They sound exactly the same |

as feature extraction for speech synthesis should preserve the personal qualities of the voice.

In this work, STRAIGHT vocoding [17], mel weighting, and cepstral analysis form the basis of the feature extraction process. STRAIGHT analysis smoothens the spectra based on knowledge of the fundamental frequency of the speech segments and so is dependant on accurate F0 estimation.

To analyse how STRAIGHT-based feature extraction for speech synthesis behaves in noisy environments, both clean and noisy data were first analysed, then resynthesised and analysed again. This distorts the waveforms quite extremely and reveals which of the vocoder parameters are most prone to voice. The results for resynthesis with noisy MCEP and band-aperiodicity components are shown in the middle columns of Table 2.

A resynthesis using clean MCEP components and noisy aperiodicity components gave fwS measures between 14.15 and 14.77 and MCD range between 1.14 and 1.18. Comparing these to the noisy resynthesis measures, it is apparent that aperiodicity components are either very robust to noise or they are far less significant in the resynthesis procedure. Either way, it is the MCEP components that suffer more from noise, and efforts to improve the resynthesis quality should focus on these.

### 3.3. Model adaptation

In a speaker adaptive speech synthesis system, an average voice is adapted to impersonate a particular speaker, i.e. to create an individual voice. The success of the adaptation is heavily dependant on the amount of available adaptation data. Preliminary experiments showed that the improvement normally attained by using multiple transforms [10] does not apply in the case of noisy data. By default all adapted synthetic voices in this work use a single global transform. Multiple transforms are used only in the case of clean and machine gun -corrupted data in one of the listening test tasks.

Results in the rightmost columns in Table 2 show that the quality of synthetic speech is significantly lower than that of the natural speech except in the noisiest of conditions. However, a comparison with the resynthesised speech indicates that the adaptation of synthetic models actually dampens the noise. In the noisiest conditions (SNR=5) the utterances synthesised from the models are better quality than the directly resynthesised utterances.

Both objective measures are compressed into a much smaller range. MCD seems to react stronger to analysis-resynthesis artifacts than HMM synthesis artifacts, and for any noise with SNR$\leq$10 indicates improvement over resynthesised speech.

## 4. LISTENING EXPERIMENT

### 4.1. Test setup

A listening test was conducted on-line, with all its benefits and troubles [18]. 26 native Finnish listeners evaluated utterances from one male and one female speaker.[1] Results from 4 listeners who did not complete the test or displayed lack of effort were discarded.

The web interface presented one question at the time in both English and Finnish, with an image showing people with mobile phones or talking on-screen to motivate the listeners to consider realistic applications. The questions used are those shown in Table 1. Each question was repeated for each sample type for both test speakers.

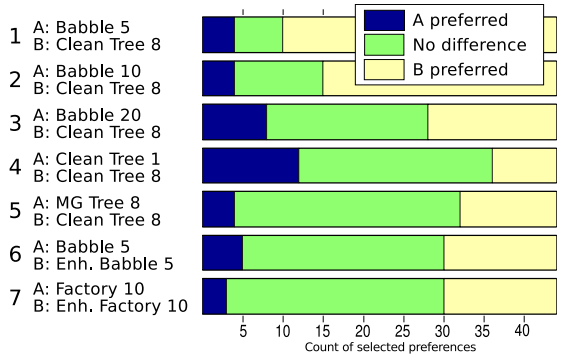
With a single training round for each question, the test consisted of 84 trials divided into two tasks. The first task was an ABX evaluation of synthetic samples listed in Figure 1 using question d from Table 1. The second task was evaluation of speaker similarity, naturalness and background quality, with natural and synthetic samples listed in Figure 2 using questions a, b and c from Table 1. The same Figure also lists the objective evaluation results for the two test speakers.

### 4.2. Results and analysis

The ABX task clearly shows that listeners can discern between adaptation from noisy and clean data, and prefer the clean one. Also,

[1] Natural and synthetic samples from these speakers can be found at http://research.ics.aalto.fi/speech/demos/noisy_synthesis_icassp13/

**Fig. 1**: Results of the listening test task 1, showing the amount of listeners preferring each choice. All samples are synthetic.




there is a slight preference to using speech enhancement techniques to remove noise.

In the second task, listeners evaluated the samples based on three different criteria (questions (a)-(c) in Table 1. The results in Figure 2 show that **naturalness** does not depend on the adaptation but all synthetic samples are considered equally (un)natural. **Similarity** to a reference sample, on the other hand, improves with speaker-based adaptation. Noisy adaptation data does not affect naturalness or similarity of the synthetic speech samples. The **background quality** is affected by noise, but the effect of the noise is less intrusive than in natural samples. This attests for the built-in noise robustness in CSMAPLR adaptation.

The evaluation of natural sentences shows that while speech enhancement does improve the perceived background quality for SNR 5 babble, it has a degrading effect on naturalness and similarity. In the case of synthetic speech, the effects on naturalness and similarity are not noticeable, whereas the background is improved.

Finally, the top rows in Figure 2 show the average fwS and MCD scores for the test speakers. While both measures reflect all three qualities evaluated in the subjective test, MCD appears to overemphasise the background qualities. That is, MCD ranks natural speech with SNR 5 babble noise lower than any of the synthetic samples. The fwS measure also depending on the background quality but shows a consistent preference to natural over synthetic samples.

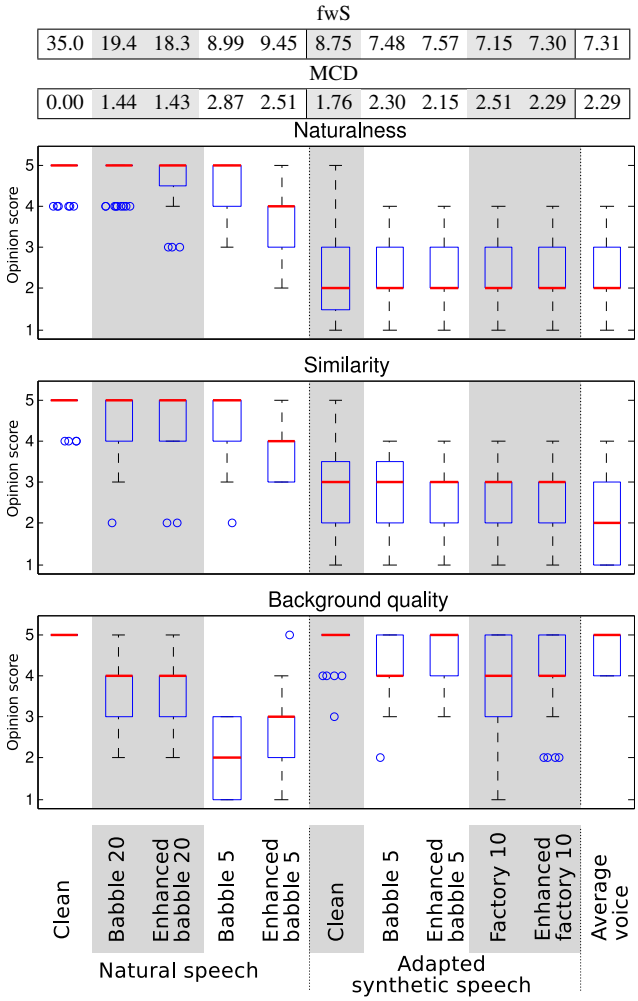
## 5. CONCLUSION

We analysed the effects of noisy environments to speech synthesis adaptation. We discovered that the analysed noises play a relatively small but noticeable role in the quality of synthetic speech, and that for practical applications, listeners prefer synthetic voices adapted from clean speech. Noise removal, even when it degrades natural speech quality, improves the synthetic voice.

We have shown that evaluation methods inspired by speech quality testing in telecommunication also reveal valuable information about the quality of speech synthesis in the presence of noise. Both objective measures, fwS and MCD, react strongly to the background improvement, but fwS correlates better with the listening test results. The proposed listening test procedure based on ITU-T 835 focuses the listener to evaluate similarity, naturalness and background qualities one at a time to give a clear picture on how the synthesis has succeeded in these dimensions and what needs to be improved.

This has been just a scratch at the surface on the issue of noise in

**Fig. 2**: Results of the listening test task 2 with natural and synthetic samples. Red bar denotes median, box extends to 25th and 75th percentiles and whiskers cover all data not considered as outliers.




HMM-synthesis, and numerous noise-removal tools and techniques in preprocessing, feature extraction and model adaptation remain to be tested. The evaluation procedure presented in this work encourages these developments and gives tools to test and measure their success.

## 6. ACKNOWLEDGEMENTS

This work received financial support from the Academy of Finland under the grants no 135003, 136209, 140969 and 251170, from Tekes under the FUNESOMO and Perso projects and from EC FP7 under grant agreement 287678. We acknowledge the computational resources provided by Aalto Science-IT project.

## 7. REFERENCES

[1] H. Zen, K. Tokuda, and A.W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, Y. Guan, R. Hu, K. Oura, Y. J. Wu, K. Tokuda, R. Karhila, , and M. Kurimo, "Thousands of voices for HMM-based speech synthesis-analysis and application of TTS systems built on various ASR corpora," *IEEE Audio, Speech, & Language Processing*, vol. 18, no. 5, pp. 984–1004, 2010.

[3] J. Yamagishi and T. Kobayashi, "Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training," *IEICE - Trans. Inf. Syst.*, vol. E90-D, no. 2, pp. 533–543, 2007.

[4] M. Aylett and J. Yamagishi, "Combining statistical parametric speech synthesis and unit-selection for automatic voice cloning," in *Proc. LangTech*, 2008.

[5] ITU-T, *Recommendation P.835 (2003/11) Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*.

[6] B. Raj, T. Virtanen, S. Chaudhuri, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *Proc. Interspeech*, 2010.

[7] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS system for Blizzard Challenge," in *Proc. Blizzard Challenge*, 2010.

[8] M. Wester, "The EMIME Bilingual Database," Tech. Rep. EDI-INF-RR-1388, The University of Edinburgh, 2010.

[9] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: Ii. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.

[10] 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," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 17, no. 1, pp. 66–83, 2009.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[12] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 229 –238, jan. 2008.

[13] J.H.L. Hansen and B.L. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Fifth International Conference on Spoken Language Processing*, 1998, vol. 7, pp. 2819–2822.

[14] 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*.

[15] M. Brookes, "VOICEBOX: Speech Processing Toolbox for MATLAB," `http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html`.

[16] M. Wester and R. Karhila, "Speaker similarity evaluation of foreign-accented speech synthesis using HMM-based speaker adaptation," in *Proc. ICASSP*, 2011.

[17] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[18] C.L. Bennett, "Large scale evaluation of corpus-based synthesizers: Results and lessons from the Blizzard Challenge 2005," in *Proceedings of Interspeech*, 2005, vol. 2005.