Noise in HMM-based speech synthesis adaptation: Analysis, evaluation methods and experiments

Reima Karhila*, *Member, IEEE,* Ulpu Remes, and Mikko Kurimo, *Senior Member, IEEE*

***Abstract*—This work describes experiments on using noisy adaptation data to create personalised voices with HMM-based speech synthesis. We investigate how environmental noise affects feature extraction and CSMAPLR and EMLLR adaptation. We investigate effects of regression trees and data quantity and test noise-robust feature streams for alignment and NMF-based source separation as preprocessing. The adaptation performance is evaluated using a listening test developed for noisy synthesised speech. The evaluation shows that speaker-adaptive HMM-TTS system is robust to moderate environmental noise.**

***Index Terms*—Speech synthesis, adaptation, evaluation methods, noise robustness.**

I. Introduction

PERSONALISED speech synthesis systems aim to mimic the voice of a specific speaker. In hidden Markov model (HMM) based speech synthesis [1], a personalised text-to-speech (TTS) system can be created based on an average voice model with speaker adaptation techniques. Speaker adaptation techniques are attractive for personalised TTS because a high-quality average voice model can be trained on data collected from several speakers and then adapted to a new speaker with just a few minutes of speech [2], [3]. Furthermore, speaker-adaptive HMM-TTS systems have demonstrated robustness to quality variations in recording conditions when the adaptation data is collected from varying sources [4], [5]. The general robustness of HMM-based speech synthesis framework has been demonstrated in [5], where condition-adaptive training was proposed to further improve the robustness towards variation in recording conditions.

Robustness to variation in recording conditions is important when the adaptation data is recorded in conditions far from studio level quality. This is typical when the samples available for speaker adaptation are found data, i.e. archived speech material recorded for purpose other than speech synthesis. An additional problem with found data or data acquired through mobile applications is background noise. The data selection process in [4] included removing recordings with background noise such as music or applause. This approach works when a sufficient amount of clean data is available, but when all samples contain background noise or the number of samples is altogether very limited, we need to cope with the noise in adaptation data.

There are two prime concerns when working with noisy data. First, personal qualities of speech can be masked or distorted by noise. To what extent information is lost depends on the noise robustness of the acoustic features. The features used in speech synthesis have been designed to capture the personal qualities of each speaker and cannot be replaced with standard noise-robust features. In principle, speech enhancement can be used to remove noise prior to feature extraction, but speech enhancement can also distort the speech signal and speaker characteristics. The other concern is whether adaptation learns the noise from adaptation data and noise is then reproduced in the synthesised speech. To what extent noise is transferred from adaptation data to synthesised speech depends on the adaptation method and parameters used in HMM-TTS adaptation.

In this work, we analyse how environmental noise affects the STRAIGHT-based acoustic features used in HMM-TTS systems and how noise transference depends on the method and parameters used in HMM-TTS adaptation. The experiments reported in this work extend our previous work [6], where we showed that HMM-TTS adaptation can learn target speaker qualities from adaptation data corrupted with environmental noise. The feature extraction and speech enhancement experiments reported in [6] are presented in more detail and the analysis is enriched with additional experiments. We then proceed to analyse how the amount of adaptation data and the number of adaptation model parameters affect noise transference and synthesised speech quality in different noise conditions. We compare HMM-TTS adaptation with constrained structural maximum a posteriori linear regression (CSMAPLR) [7] and eigenspace-based maximum likelihood linear regression (EMLLR) [8], and investigate using a noise-robust feature stream for model alignment.

We note that standard evaluation methods used with HMM-TTS adaptation are not reliable when noisy adaptation data is used. The standard objective measures and listening tests have been developed to evaluate how well a synthesised voice corresponds to a target speaker in clean background, and do not take into account noise transferred from the adaptation data. Speech and noise are perceptually different and each listener weights the background noise intrusiveness and the perceived speech distortion according to their personal preference. This introduces excess variation in listening test results and reduces the reliability of the test as an indicator for the best system [9].

Manuscript received April 30 2013; revised July 10 2013. This work was supported in part by the Academy of Finland in projects 135003, 140969, 251170, Tekes in Perso and FuNeSoMo project, and EC FP7 under grant agreement 287678. R. Karhila was supported by Langnet graduate school and Nokia Foundation.

R. Karhila, U. Remes, and M. Kurimo are with the Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, Finland (e-mail: reima.karhila@aalto.fi)

In this work, we evaluate HMM-TTS adaptation quality using the objective measures and listening test procedure proposed in [6].

The remainder of this work is organised as follows. Section II describes the features and methods used in this work. Section III introduces the objective measures and the subjective listening test setup used for evaluation and Section IV describes the evaluation data and the system parameters used in this work. Experimental results that concern feature extraction are presented in Section V and results that concern model adaptation in Section VI. The feature extraction and adaptation results are discussed in Section VII.

II. Methods

A. Feature extraction

Features set an upper limit to adaptation performance since adaptation methods cannot learn speaker characteristics not represented in the features. While numerous feature extraction methods exist for improving noise robustness in automatic speech recognition, for example, many of these techniques normalise the speech signal and cannot be applied to personalised TTS.

In [10], a single-channel non-negative matrix factorisation (NMF) based speech enhancement method proposed in [11] was succesfully applied to remove background noise from average model training data in a HMM-TTS task. The results on average voice model training do not indicate if the enhanced speech signal retains the personal characteristics of a speaker. In this work, we review and extend the experiments reported in [6] where the speech enhancement method [11] was applied on HMM-TTS adaptation data.

B. Model alignment in adaptation

In HMM-TTS adaptation, each average voice model state is modified to mimic the characteristics of a specific speaker. Adaptation performance depends on proper model alignment because the state alignments to determine which adaptation data samples are used to update a specific average model state. When the adaptation data contains noise, there is a mismatch between the average voice models and the observed features. Noise thus increases the errors in the state alignment sequence and limits the quality that can be achieved with HMM-TTS adaptation.

We investigate two approaches to improve noise robustness in model alignment. (1) State boundaries can be estimated using a noise-robust feature set, after which transformations are computed from the actual TTS features. (2) Transformations can be computed iteratively, starting from a simple global transformation which will be used to get improved state alignments to calculate the final adaptation transformations.

C. CSMAPLR adaptation

The most common adaptation methods used in both ASR and TTS are based on linear transformation matrices applied to model parameters. The linear transformation methods for model mean adaptation include maximum likelihood linear regression (MLLR) and maximum a posteriori linear regression (MAPLR) methods, and methods for joint mean and covariance transformation include the constrained MLLR (CMLLR) [12]. When enough adaptation data is available, detailed transformation matrices can be trained for clustered phone models. The number of transformations, and thus, the level of detail and personality transferred from the adaptation data to the adapted models can be controlled with decision trees. The constrained structural MAPLR (CSMAPLR) method [7] combines joint mean and variance adaptation with recursive MAP-based parameter estimation in the decision tree framework.

MLLR or CMLLR transformations are calculated based on statistics accumulated from the adaptation data, and are robust to occasional noises when enough data is available to accumulate the statistics for each transformation. The accumulators are state-dependent, so the adaptation data must be segmented and aligned. CSMAPLR uses soft alignments i.e. estimated model occupancy probabilities. The method requires around 3–6 minutes of clean speech data to generate high-quality personalised voices [2], but around 1 minute is sufficient to increase the perceived speaker similarity [13]. When noisy adaptation data is used, significantly larger sets may be needed for comparable performance, and also the regression tree size needs to be controlled to minimise noise transference. Noise transference naturally depends on the noise type, and short noise bursts like gunshot noise are efficiently averaged out in adaptation [6].

D. Eigenvoice-based adaptation

Eigenvoice adaptation [14] is a robust method for speaker-based mean adaptation based on limited amounts of adaptation data. The method assumes N speaker-dependent acoustic model sets constructed from training data. In standard eigenvoice adaptation, the mean vectors in each model set n are concatenated into a supervector $\tilde{\boldsymbol{\mu}}_n$, and principal component analysis (PCA) is applied on the N speaker-dependent supervectors to calculate M eigenvectors $\boldsymbol{v}_1 \dots \boldsymbol{v}_M$, where $M << N$. The adapted model means $\tilde{\boldsymbol{\mu}}_i$ for a target speaker i are then calculated as a linear combination of the eigenvectors,

$$\tilde{\boldsymbol{\mu}}_i = \sum_m w_{i,m} \boldsymbol{v}_m, \tag{1}$$

where $w_{i,m}$ are scalar weights estimated from the adaptation data [14].

The standard eigenvoice adaptation is a relatively common technique in ASR, and has been succesfully applied to speech synthesis [15]. Since the adapted model means are calculated as linear combinations of a limited number of basis vectors, eigenvoice adaptation typically looses some personal speaker qualities, but is also not expected to transfer distortions such as noise from the adaptation data to the adapted voice model; eigenvoice adaptation resembles PCA-based methods used for image denoising [16].

Eigenvoice adaptation does not require much data from the target speaker since parameters estimated based on adaptation data include only the M scalar weights. Training data for the N speaker-dependent models, on the other hand, needs to be

pooled from several speakers in order for the eigenvectors to be representative for a variety of speakers. The requirements for training data in eigenvoice adaptation are difficult to meet especially in HMM-TTS systems, where the acoustic models are typically larger than acoustic models used in ASR and the requirements for training data quality are strict. Furthermore, supervectors constructed from such large model sets would be too high-dimensional for the standard principal component analysis.

In this work, PCA is applied on speaker-dependent MLLR transformations, as proposed in [8], rather than speaker-dependent model sets. The eigenspace-based maximum likelihood linear regression (EMLLR) adaptation [8] proceeds as follows:

1) Train an average voice model and estimate linear transformations to adapt the Gaussian means for N training speakers. Note that CMLLR transformations cannot be used as EMLLR does not handle covariance transformation.
2) Reshape the N speaker-dependent transformation matrices into supervectors, normalise with the mean and covariance, and compute the eigenvectors associated with the M largest principal components of the supervectors.
3) Denormalise and reshape the eigenvectors into MLLR transformation matrices $\boldsymbol{V}_m$.
4) The acoustic model means $\boldsymbol{\mu}_i^{(g)}$ for target speaker i are calculated as

$$\boldsymbol{\mu}_i^{(g)} = \sum_m w_{i,m} \left(\boldsymbol{V}_m^{\mathsf{T}} \boldsymbol{\xi}^{(g)} \right), \tag{2}$$

where $\boldsymbol{\xi}^{(g)} = [\boldsymbol{\mu}^{(g)\mathsf{T}} 1]^{\mathsf{T}}$ is the augmented mean vector of the Gaussian in acoustic model state g and $w_{i,m}$ are the speaker-dependent weights estimated from adaptation data.

EMLLR combines the robustness of eigenvoice methods with the average voice based adaptation paradigm introduced in [2]. This allows using large databases not specifically intended for speech synthesis training to generate the N speaker-dependent transformations.

A technique similar to the eigenvoice is cluster-adaptive training (CAT)[17] where speaker adaptation can be performed rapidly using by combining clusters of Gaussian mean parameters learned during average model training. It has also been applied to TTS [18], [19]. Eigenvoice adaptation can be seen as a simplified version of CAT, and can actually be used to initialise CAT parameters.

Finally, we note that kernel methods can be used to replace the standard PCA, and the eigenvectors can be calculated in a mapped feature space to overcome problems related to high-dimensional supervectors [20], [21], [22]. The kernel methods allow for nonlinear reconstruction of the target voice models, and can also be used to improve the EMLLR performance in recognition tasks, as proposed in [23].

III. Evaluation methods

A. Objective evaluation methods

The common measures for objective evaluation of speech synthesis quality include mel-cepstral distortion calculated between natural speech sentences and the corresponding synthesised sentences. Mel-cepstral distortion [24] is calculated for D-dimensional features as

$$MCD = \frac{1}{M} \sum_{m=1}^{M} \sqrt{2 \sum_{d=0}^{D-1} (c(d,m) - \hat{c}(d,m))^2}, \tag{3}$$

where $\hat{c}(d,m)$ and $c(d,m)$ denote the dth coeffcient in test and reference mel-cepstra in time frame m, and M denotes the number of frames.

In this work, mel-cepstral distortion is used in conjunction with a perceptually motivated distortion measure that is more reliable in noisy conditions. The measure used in this work is based on frequency-weighted segmental SNR (fwS) [25]. The measure has been shown to correlate well with the industry standard objective evaluation method PESQ [26], but is substantially cheaper to implementat and compute [27]. The measure is calculated as

$$fwS = \frac{10}{M} \sum_{m=1}^{M} \frac{\sum_{j=1}^{K} W(j,m) \log_{10} \frac{X(j,m)^2}{((X(j,m) - \hat{X}(j,m))^2}}{\sum_{j=1}^{K} W(j,m)}, \tag{4}$$

where $\hat{X}(j,m)$ is the test signal value in the jth mel filter channel in time frame m, $X(j,m)$ is the reference signal value in the same mel channel, and $W(j,m) = X(j,m)^{\gamma}$ with $\gamma = 0.2$ as proposed in [27]. The test and reference signals were processed in 25 ms windows with 5 ms shift between adjacent frames, and represented with FFT spectra. The $K = 21$ channel mel-filterbank used to calculate mel-spectral features was estimated with VOICEBOX [28]. The estimated SNR in each time frame was bound to $[0, 35]$ dB as in [27].

To use the objective measures to compare a reference sample and a sample generated with the HMM-TTS system, the synthetic sample is generated based on the phone alignment of the reference sample. The objective measures were then calculated based on 2 second samples extracted from the middle of the utterances. Since the synthesis system may introduce excess frames in the sample, the comparison between the reference and test samples was done at a varying frame delay ($[-10, ..., 10]$) and the best match was reported.

B. Subjective evaluation methods

In the widely used subjective evaluation method of mean opinion scores (MOS) known from the Blizzard challenge [29], listeners listen to natural and synthesised speech samples, and judge them on a subjective scale from one to five regarding both naturalness and similarity. When noisy training data is used for HMM-TTS adaptation, some background noise can be transferred to the synthesised samples, and the standard test results can become unrealible. A standard MOS test with synthetic, noisy samples indicated that listeners have widely varying reactions to noise, and thus, it was not possible to find significant average differences even between strongly differing samples. In particular, the results from the MOS test suggested that background noise can mask synthesis artifacts, as the listeners gave more favourable judgements to the noisier samples.

In this work, we use the subjective listening test proposed in [6]. The test procedure is modelled after the principles used in testing telecommunication systems with noise suppression algorithms [9], where listeners hear the same sample three times with instructions to attend different qualities in the sample, namely how distorted the speech signal is, how noisy is the background signal and how the listener finds the overall quality.

The test framework used in this work has two parts. The first part is an AB test aimed to survey which systems would be preferred for everyday use. The distinction between the speech signal and background is not emphasised in this part. The second part is the modified MOS test where listeners are asked to rate either the speech signal or the background quality. The test setup and instructions proposed for the evaluation of speech synthesis systems adapted using noisy data are presented in Table I.

The HMM-TTS adaptation procedures used in this work affect also the silence models, which means adaptation can introduce synthesised noise to the beginning and end of the synthesised utterances. Since this may allow the listeners to adapt to the background noise and affect evaluation, the ITU-T listening test procedure for noisy channels [9] suggests that listeners hear only 4 second samples of speech, not complete utterances, with silent breaks at the beginning and the end. Since the samples used in this work were generally shorter than 4 seconds, the whole utterance parts were presented to the listeners but the silence in the beginning and end were cut off.

TABLE I
QUESTIONS USED IN THE SUBJECTIVE EVALUATION

▷ Natural reference speech sample
▷ Synthesised speech sample A
▷ Synthesised speech sample B

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.
Regarding the OVERALL QUALITY

A. First sample is better
B. Second sample is better
C. They sound exactly the same

▷ Synthesised speech sample

Play the sample and attending ONLY to the SPEECH SIGNAL, select the category which best describes the sample you just heard.
the SPEECH SIGNAL in this signal was

5. Completely natural
4. Quite natural
3. Somewhat unnatural but acceptable
2. Quite unnatural
1. Completely unnatural

▷ Synthesised speech sample

Play the sample and attending ONLY to the BACKGROUND, select the category which best describes the sample you just heard.
the BACKGROUND in this signal was

5. Clean
4. Quite clean
3. Somewhat noisy but not intrusive
2. Quite noisy and somewhat intrusive
1. Very noisy and very intrusive

▷ Natural reference speech sample
▷ Synthesised speech sample

Play both samples, and attending ONLY to the SPEECH SIGNAL, select the category which best describes the second sample to the reference sample.
The voices in the SPEECH SIGNALS of the samples sounded

5. Exactly like the same person
4. Quite like the same person
3. Somewhat different but recognisable
2. Quite like a different person
1. Like a totally different person

IV. Experimental setup

A. Evaluation data

The experiments reported in this work were conducted on the Finnish utterances from the EMIME TTS corpus [30]. The utterances are short sentences recorded from native Finnish speakers in a studio environment with close-talk microphone. The data used in this work contains 145 utterances from three male and three female speakers. Depending on speaker, this amounts to 6-11 minutes of audio data, of which 5-7 minutes are non-silent segments for an average 2-2.8s of speech per sentence. To evaluate HMM-based speech synthesis adaptation, the data is divided into separate adaptation and test datasets of 105 and 40 sentences respectively.

To evaluate how noise affects HMM-based speech synthesis and speaker-based adaptation, the utterances from the EMIME dataset were mixed with samples from the NOISEX-92 [31]. The noises used in this work include babble noise, factory noise, and machine gun noise that were mixed with the clean speech utterances at several signal-to-noise ratios (SNR). The noise corrupted utterances were created as follows. For each utterance, a noise sample of the same length as the clean speech utterance was randomly extracted from the noise recording. The average energy of the clean speech and noise samples were calculated and the noise sample was scaled to adjust the SNR. Figure 1 shows scaled STRAIGHT and FFT spectra of a samples corrupted with different background noise types.

B. Average voice models

Male and female average voice models were trained using the methods and tools of the EMIME 2010 Blizzard Entry [32]. The models are context-dependent multi-space distribution hidden semi-Markov models (MSD-HSMM) trained on the new Finnish PERSO corpus[1]. The training data for the male average voice model contains 4200 utterances (310 min; 220 min of speech) from 20 speakers and the training data for the female average voice model 3900 utterances (250 min; 155 min of speech) from 30 speakers. The utterances are read sentences recorded with a close-talk microphone in a studio

[1]The data will be publicly released soon.

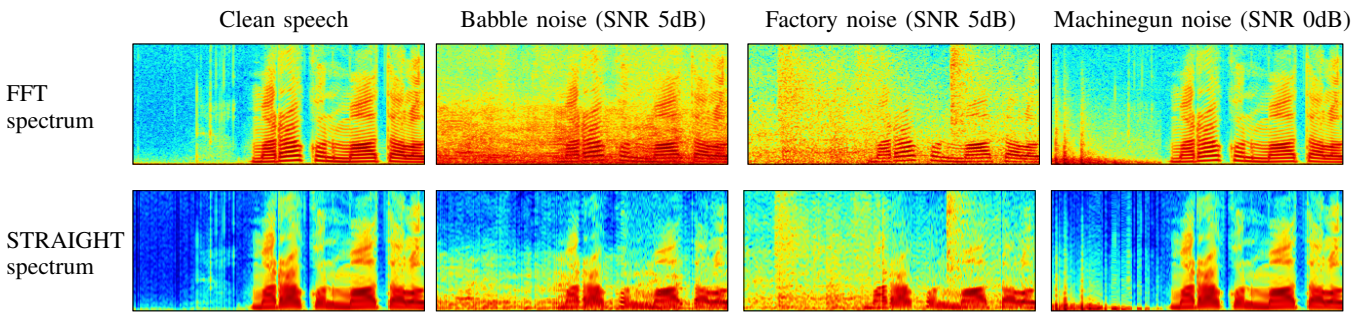


Fig. 1. Log-STRAIGHT spectra and Log-FFT spectra of a sample training sentence from female speaker in various noise conditions. First half of the sample is silence (or pure noise), second half speech.

environment. Speaker-adaptive training was applied to create a speaker-adaptive average voice model.

The average voice models are trained on acoustic feature vectors calculated based on STRAIGHT vocoder [33] output. The feature streams used in the model include STRAIGHT-analysed mel-generalised cepstral coefficients (MCEP), band-limited aperiodicity (BNDAP) components, and fundamental frequency (F0). The F0 feature used in STRAIGHT analysis is calculated as the median of F0 estimates calculated with three methods: IFAS [34], fixed-point analysis (TEMPO) [35] and ESPS [36].

In this work, the acoustic feature vectors were augmented with the ETSI advanced front-end (ETSI-AFE) features [37]. These are noise-robust features that have been developed for automatic speech recognition in noisy environments. The features are based on mel-frequency cepstral coefficients (MFCC) derived from Wiener-filtered data, and were included in this work to investigate if using a noise-robust feature stream improves model alignment during adaptation computation. The ETSI-AFE features were not used in training the average voice model.

C. Speech enhancement

The NMF-based source separation method proposed in [11] models the noisy speech features as a linear combination of speech and noise exemplars. The speech and noise codebooks used in this work were generated as follows. The exemplars in the clean speech codebook are random samples from the average model training data and additional clean speech utterances from 309 speakers from the Finnish SPEECON database [38]. The SPEECON utterances are read sentences recorded with a close-talk microphone in quiet office environments. The exemplars in the noise codebook are random samples from the NOISEX-92 data recorded in babble and factory noise environments. We note that the clean speech and noise samples used in codebook generation do not overlap with the samples used in the evaluation data.

The method was tuned to the speech synthesis task using part of the EMIME corpus data. A computationally expensive grid search was used to find the parameters that minimise the average MCEP distortion between the enhanced speech samples and the original clean speech samples. Later, it was found out that a small portion of tuning data overlapped with the evaluation data. However, we think that this had negligible effect on the performance.

D. Eigenvoice adaptation

Eigenspace-based adaptation requires a large collection of speaker-dependent voice models or MLLR transformations. In this work, the male and female average voice models (Section IV-B) were used as baseline model sets and global block-diagonal MLLR transformations were computed for all the training speakers in PERSO corpus. MLLR transformations were additionally computed for 163 male and 146 female speakers from the SPEECON corpus [38]. The adaptation datasets from SPEECON had around 3–4 minutes of speech per training speaker. We note that this would not have been sufficient for training personalised model sets for standard eigenvoice adaptation, and thus, EMLLR was chosen for this work.

A separate eigentransformation set was computed for the male and female average voice models, but all training speakers (male and female) were adapted to both models to ensure a large enough transformation base, $N = 359$. The eigenvectors of the MLLR transformation sets for were solved with singular value decomposition and eigenvectors corresponding to the $M = 17$ largest principal components were included in the eigentransformation set $\{\boldsymbol{V}_1, \ldots \boldsymbol{V}_M\}$.

V. Experiment 1: Features

A. Noise in natural speech

To obtain a reference level for noise distortion, we evaluate the effect of noise in the adaptation training data represented with standard FFT and mel-cepstral features. The first columns in Table II present the objective evaluation measures calculated for the adaptation data under different noise conditions. Distortion by noise is indicated as deviation from the optimal score (35 dB and MCD 0) in top row. The objective measures indicate that factory noise is more intrusive than babble noise, and that the continuous noises (babble and factory) degrade the overall speech quality more than the occasional burst of noises (machine gun).

When the noisy adaptation data is preprocessed with speech enhancement, the results improve in the noisy cases, SNR $\leq$

TABLE II
AVERAGED FWS AND MCEP DISTORTION MEASURES FOR 3 MALE AND 3 FEMALE SPEAKERS IN VARIOUS NOISE CONDITIONS FOR (A) NATURAL SPEECH, (B) VOCODER-ANALYSED AND RESYNTHESISED SPEECH AND (C) SYNTHETIC SPEECH GENERATED WITH CSMAPLR-ADAPTED HMMS.

		(A) Original training data		(B) Vocoder-resynthesised training data		(C) adapted HTS-synthesised test data	
Noise	SNR	fwS	MCD	fwS	MCD	fwS	MCD
Clean	-	35.0	0	15.0	1.1	9.6	1.9
Babble	20	19.7	1.3	13.2	1.9	8.9	2.1
	10	12.4	2.2	10.3	2.9	8.5	2.4
	5	9.1	2.7	8.3	3.4	8.1	2.6
Factory	10	9.5	2.9	8.4	3.6	7.8	2.7
	5	6.6	3.4	6.2	4.2	7.3	3.0
Machine Gun	0	20.7	1.1	12.9	1.7	8.9	2.0
Enhanced Babble	20	18.6	1.4	13.3	1.9	8.9	2.0
	10	12.7	2.0	10.6	2.6	8.5	2.2
	5	9.7	2.2	8.9	3.1	8.2	2.4
Enhanced Factory	10	10.6	2.2	9.2	3.0	7.8	2.4
	5	8.2	2.4	7.6	3.4	7.5	2.6

10 dB, but when the noise in the original data is less intrusive, the distortion introduced in the speech signal starts to balance with the distortion (noise) removed.

B. Noise in vocoder

As the parametric speech synthesis system relies on estimating models for specific components of a source-filter sound production system, the used vocoder sets the limit for the naturalness of the generated speech. The STRAIGHT-based feature set used in this work includes three parameter streams. The STRAIGHT-MCEP feature stream determines the filter parameters and the aperiodicity and F0 features affect the excitation model. To analyse how noise and speech enhancement affect the STRAIGHT-based features, we compare features calculated from the clean speech data and the noise-corrupted or enhanced data. The data is resynthesised based on the clean or noisy features, and the resynthesised samples are subjected to objective evalution.

1) MCEP components: The objective evaluation results for samples resynthesised from noisy MCEP and band-limited aperiodicity (BNDAP) components are presented in columns (B) in Table II. We see that the range of scores is compressed compared to scores of natural speech in columns (A). On the top row, we have the absolute limit for synthesis quality for this particular dataset: fwS 15.0 dB and MCD 1.1. The following rows show the degradation as measured with our objective measures. As with natural speech, the quality degradation is steepest when factory noise is used. Speech enhancement again improves the results when SNR $\leq$ 10 dB.

2) F0 component: STRAIGHT analysis smoothens the spectra using knowledge of the fundamental frequency of the speech segments and is therefore dependant on accurate F0 estimation. To analyse the error introduced by noise to F0 extraction, the F0 features calculated from the samples in adaptation data were compared with the F0 extracted from the original clean speech samples to calculate a mean error in the voiced frames. The mean error and the percentage of voicing errors are reported in Table III. The results indicate that all noise types investigated in this work introduce error in the F0 extraction. Using speech enhancement reduces the amount of voicing errors in factory noise corrupted data, but increases the mean error in voiced frames in both babble and factory noise data.

TABLE III
F0 EXTRACTION ERRORS FOR 3 MALE AND 3 FEMALE SPEAKERS IN DIFFERENT NOISE CONDITIONS, AND THE FWS VALUES OF TRAINING DATA RESYNTHESISED USING F0 EXTRACTED FROM THE NOISY DATA AND MCEP AND BNDAP COMPONENTS FROM THE CLEAN DATA.

Noise type	SNR	Mean error in voiced frames (Hz)	Voiced/unv. error (% of frames)	fwS of resynth. speech
Babble	20	1.4	2.7	15.0
	10	2.3	7.4	14.6
	5	4.2	13.7	14.1
Factory	10	1.7	6.5	14.7
	5	2.0	12.9	14.0
Machinegun	0	8.7	12.4	14.0
Enhanced Babble	20	2.2	3.7	15.0
	10	3.5	8.5	14.9
	5	6.5	13.5	14.5
Enhanced Factory	10	3.2	5.1	14.9
	5	4.8	8.0	14.6

Finally, the rightmost column in Table III presents fwS scores calculated for the data that has been analysed and resynthesised using F0 extracted from noisy data and MCEP and BNDAP components extracted from clean speech data. The results suggest that errors in F0 extraction translate into quality degradation in the resynthesised signals, but the effect is smaller than the effect from using noisy MCEP and aperiodicity components (Table II columns (B)). The scores for noisy data are in range 14.1–14.7 dB as opposed to 6.2–13.2 dB when noisy MCEP and aperiodicity components were used.

3) Aperiodicity components: When the data was resynthesised from clean MCEP and noisy aperidocity components, the fwS scores were in range 14.2–14.8 dB and MCD scores in range 1.1–1.2. The scores are not reported here in full, but comparing even the range to the noisy resynthesis measures reported in columns (B) in Table II makes it apparent that aperiodicity components are either very robust to noise or far less significant than MCEP components in the resynthesis procedure.

STRAIGHT allows using band-limited or standard, non-limited aperiodicity for the excitation parameters. Both band-limited and standard aperiodicity were tested, and although the differences were small, we could notice that band-limited aperiodicity is more resistant to babble noise and standard aperiodicity more resistant to factory noise.

C. Quality of synthesised speech

The columns (C) in Table II show the measured performance for samples synthesised from models adapted with CSMAPLR CSMAPLR transformations for each target speaker were estimated from 105 utterances (5-7 minutes of speech depending on the speaker). Compared to resynthesised utterances, HMM-synthesis further compresses the scale, with the average fwS score being in the range 7.3-9.6 depending on the noise, compared to 6.2-15.0 of resynthesised speech. This shows the robustness of HMM-synthesis. With the addition of noise, the quality of the adapted speech decreases far less than in the case of natural or vocoder-synthesised speech. According to the objective measures, in the noisiest factory case, the adapted synthetic speech is actually of better quality than the noise-corrupted natural speech of the same type. The addition of noise in the training data muffles the voice, and loud noise introduces some extra fuzziness in the background, but the objective measures judge this to be less degradating than the loud, clear noise signal in the noise-corrupted training data.

The improvements shown by NMF-enhancement on the training data are not as dramatic in the fwS scores of the final synthetic voices. Nonetheless, there is clear improvement a for the low SNR cases.

D. Subjective evaluation

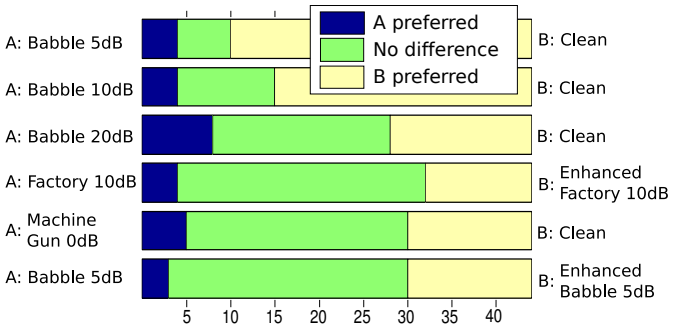
A subjective listening test was conducted to evaluate how noise and speech enhancement affect listener preferences in HMM-TTS adaptation task. The setup described in Section III-B was implemented as a web interface that presented the questions listed in Table I one question at the time. The order of the evaluated samples was shuffled for each listener based on a checksum generated from the listener's registration email address. The questions were presented both in English and Finnish. The test samples used in the evaluation represented one male and one female speaker, and the test questions listed in Table I were repeated for both test speakers for each processing type evaluated in the test.[2]

26 listeners who identified as native Finnish speakers participated in the test. The participants included 4 listeners who either did not complete the test or displayed a clear lack of effort. Their input has been discarded. The listening test results presented here were first reported in [6] and are included here for completeness.

The listening test was divided in two separate tasks. In the AB evaluation task, the listeners compared samples generated with personalised HMM-TTS systems. The adaptation data sets contained 105 utterances (5–7 minutes of speech depending on the speaker) with different noise types and noise levels and optional speech enhancement. The listener preferences are reported in Figure 2. Statistical significance tests (Binomial test, $H_o : prob(A) = prob(B) = 0.5$, "No difference" is counted as a half vote for both of the compared cases) indicate that the listeners can discern between systems adapted from noisy and clean data and prefer the system trained with clean data over data with babble noise at SNR 10dB ($p = 0.006$) and 5dB ($p = 0.001$). The listening test does not support any claims for listener's preference when comparing clean systems and systems with babble at SNR 20dB and machine gun noise at SNR 0dB ($p = 0.21$ for both). The listening tests show no strong evidence that using speech enhancement to remove noise from the adaptation data would result in higher speaker preference ($p = 0.18$ for babble noise and $p = 0.13$ for factory noise).

[2]The samples used in the test are available in http://research.ics.aalto.fi/speech/demos/noisy_synthesis_icassp13/

Fig. 2. AB test results from listening test 1. The samples have been generated with models adapted using clean, noisy, or enhanced data. The noise types in this test include babble and factory noises introduced at SNR 5–20 dB and machinegun noise at 0 dB. A single transform is used for adaptation in babble and factory noise cases. A tree of 8 transforms is used for clean and MG noise cases.



Results from the second task are illustrated in Figure 3. The samples included natural speech and samples generated with the personalised HMM-TTS systems adapted with 105 utterances. The results from the second task can be summarised as follows.

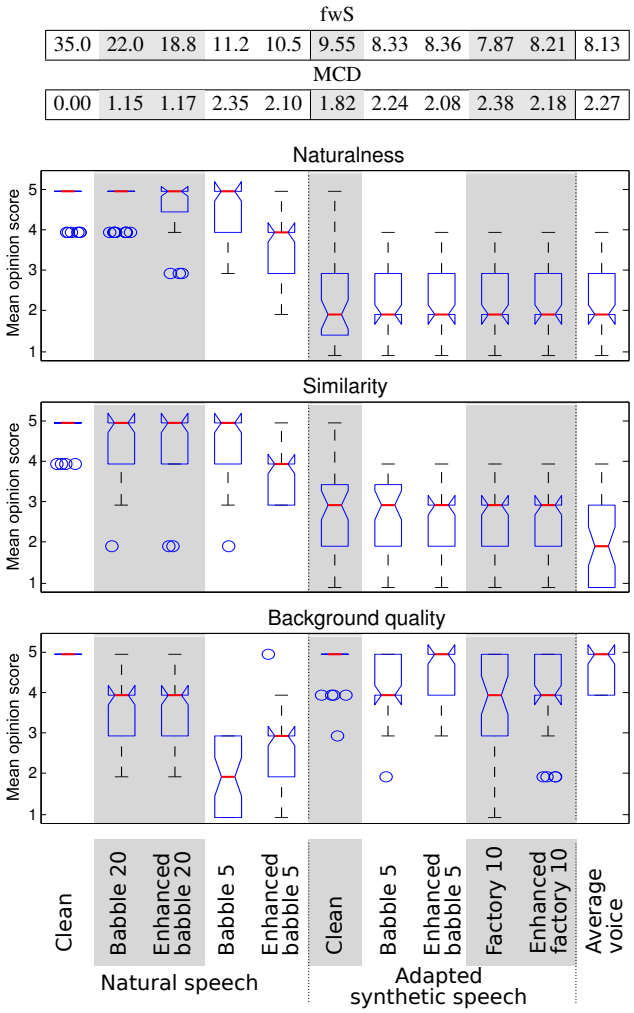
1) Naturalness: Results on the synthesised speech samples indicate that adaptation does not affect naturalness, but samples generated with the average voice model are rated same as samples generated with the adapted models. We note that the samples corrupted with babble noise at SNR 5 dB are perceived less natural when processed with speech enhancement, whereas using noisy or enhanced data for adaptation does not affect the naturalness.

2) Similarity: Results on the synthesised samples indicate that the similarity between the test samples and a reference sample (natural clean speech) improves with speaker-based adaptation. While the similarity between the reference sample and natural speech samples corrupted with babble noise at SNR 5 dB degrades when speech enhancement is used, using noisy or enhanced speech data for the HMM-TTS adaptation task does not affect the perceived similarity.

3) Background quality: The synthesised samples have some background noise when noisy adaptation data is used, but the noise is not considered as intrusive as the noise in the natural speech samples. Furthermore, while the results on natural speech samples indicate speech enhancement does not remove all the noise, using speech enhancement on the adaptation data results in synthesised samples with a clean background. The results attest for the built-in noise robustness in CSMAPLR adaptation.

The samples used in the listening test were also subjected

Fig. 3. Results of the listening test A task 2 with natural and synthetic samples speakers F2 and M3. Red bar denotes median, box extends to 25th and 75th percentiles and whiskers cover all data not considered as outliers. 95% confidence interval of the median is denoted by the notches.



to objective evaluation and the average scores in each test condition are reported in Figure 3. The measures are correlated to an extent with all three qualities evaluated in the subjective test, but as discussed in [6], MCD emphasises the background quality and ranks clean synthesised samples better than natural speech corrupted with loud babble noise. The linear correlation coefficients ρ between the reported average scores and the average MOS scores are as follows. The correlations between the fwS scores and MOS scores in naturalness, speaker similarity, and background quality are $|\rho| = 0.77$, $|\rho| = 0.77$, and $|\rho| = 0.18$, and between the MCD and MOS scores $|\rho| = 0.68$, $|\rho| = 0.67$, and $|\rho| = 0.34$.

VI. Experiment 2: Adaptation strategies

A. Model alignment

The average deviations between the noisy and clean boundaries are reported Table IV. The ST-MCEP column shows that alignment with the STRAIGHT-based features becomes difficult with the addition of noise. Alignment is also slow and requires high beams. For some utterances, even a high beam would not work, and thus, any utterances that would have required a search beam over 10000 were excluded from the analysis.

TABLE IV
Average deviation of HMM state border in noisy cases in frames. Silence models are excluded. Reference alignments are from clean speech.

Noise type	SNR	Boundary deviation (frames)			
		ST-MCEP	+Adapt	Aurora	+Adapt
Babble	20	2.6	1.6	1.1	1.3
	10	5.0	2.9	2.6	3.2
	5	11.8	5.3	6.1	6.7
Factory	10	7.9	3.1	3.6	3.6
	5	17.4	5.8	7.0	5.9

The Aurora column in Table IV shows the state boundary deviations when the ETSI-AFE noise-robust feature set is used for alignment. In all cases but one, the Aurora feature stream outperforms the STRAIGHT MCEP feature stream, and also executes faster and more reliably. When Aurora features are used, an alignment is always found with a comparatively low beam.

Finally, generating a single global transform to adapt to the noise and using that transform to improve the alignment performance with STRAIGHT MCEP features results in alignments as accurate or better than alignments generated based on the Aurora feature stream. This increases computation time, but if iterative generation of adaptation transforms is possible, the added complexity of using the Aurora feature stream is hard to justify. In the remainder of the experiments reported in this work, the state alignments are generated iteratively using a single global transformation to adapt the average voice model.

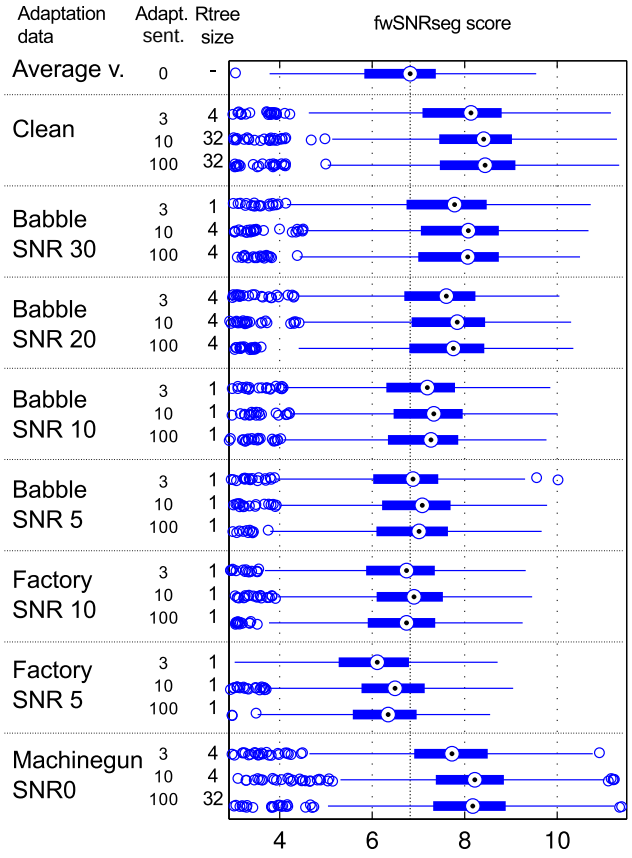
B. CSMAPLR adaptation

This experiment shows how the amount of data used for training the adaptation transforms affects the synthetic speech quality. CSMAPLR transformations were trained using subsets of the training data. This was done five times over a random subset in order to average out the differences in training qualities between sentences. The same randomly selected sentence sets were used in all noise conditions.

The results are presented in Figure 4. The objective evaluation measure indicates that increasing the number of samples does not significantly improve the synthesised speech quality when adaptation is corrupted with continuous noise like babble or factory noise. As an exception, speech degraded with machine gun noise can be used in larger quantities to diminish the gap to the samples generated with CSMAPLR transformations estimated based on clean speech.

The results in Figure 4 also demonstrate how the nature of the noise in adaptation data needs to be considered when deciding the geometry of the adaptation transforms. The adaptation experiment was repeated for regression tree sizes with 1, 4 or 32 leaves, and only the best score is reported in Figure 4. In the case of clean speech, a greater number of transformations produces better quality speech as evaluated with both MCD and fwS measurements. For any significant

Fig. 4. Frequency-weighted segmental SNR for synthesised samples for three male and three female speakers when adaptation data for CSMAPLR is corrupted with different noise types. Adapt sent. column indicates the number of utterances used for adaptation and Rtree size column the size of the regression tree. The boxes show the 25th and 75th percentiles and the circle within the box indicates the median. The lines cover all the data points included in the analysis. Data points deviating more than 2σ from the mean are marked with circles.



amount of noise, the best score is obtained with a single global transform.

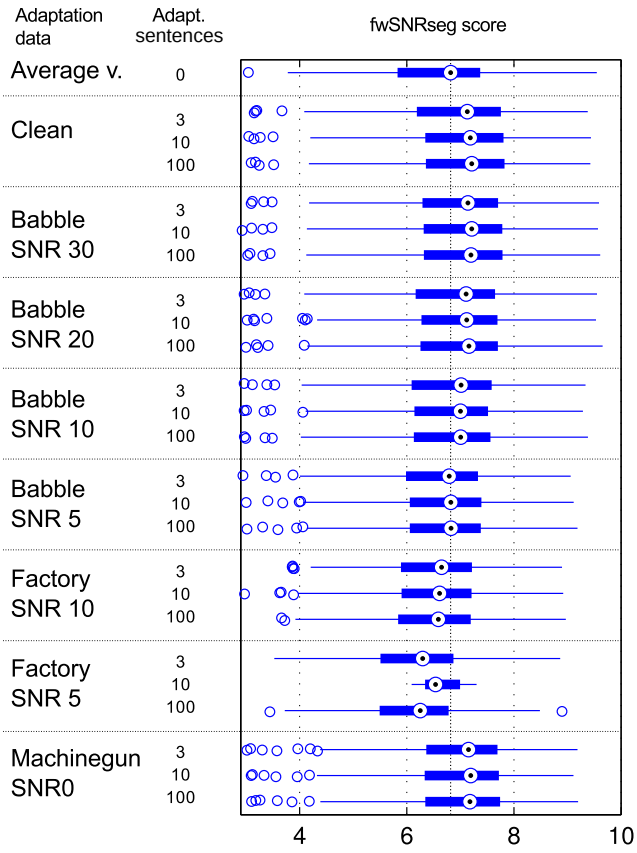
C. Eigenvoice adaptation

The results of the objective evaluation of EMLLR adaptation in various noise conditions for various amounts of adaptation data are shown in Figure 5. Very little data is required for the basic adaptation, and there is very little improvement for using more data. This is in line with the known aspects of eigenvoice systems.

The fluctuations in quality follow approximately those of the CSMAPLR adaptation, and while in the noisiest conditions (factory at SNR 5) the methods perform equally badly, in the case of clean data, CSMAPLR adaptation has a clear edge. EMLLR is not as good in capturing the fine subtleties of the target speaker, but the scores have a smaller range than CSMAPLR, which suggests some robustness to noise. At least in theory, that eigen-space based adaptation methods would be less prone to include noise in adapted model set.

We note that while EMLLR would be robust towards noise in that the final adapted is relatively noise-free, the alignments used to calculate the EMLLR transforms are worse in the

Fig. 5. Frequency-weighted segmental SNR for synthesised samples for three male and three female speakers when adaptation data for EMLLR is corrupted with different noise types. Adapt. size column indicates the number of utterances used for adaptation, and the graph is formatted as described in Figure 4.



noisy cases compared to clean, which could lead to general degradation adaptation quality and produce a less personal voice. It is notable that our EMLLR setup is based on hard state boundaries, and might therefore be more susceptible to alignment problems than the linear adaptations calculated with soft boundaries.

D. Subjective evaluation

The second subjective analysis was done to test the hypothesis that eigenspace based adaptation methods would be more robust to noise than the standard linear regression adaptation methods. The test therefore employed the noisiest data used in our experiments. Since eigenspace-based methods generally work with very small sets of adaptation data, and the adaptation data was limited to 10 utterances per target speaker. The adaptation dataset for each speaker contained around 30 seconds of speech.

31 listeners evaluated the samples in a setup identical to the subjective evaluation in Section V-D. Samples for two male and two female target speakers were generated with model sets adapted with either EMLLR or a global CSMAPLR transformation. Both methods used one CMLLR transform to help in properly aligning state boundaries as proposed in Section VI-A. Samples generated with the male and female

Fig. 6. Results of the listening test 2 AB task comparing CSMAPLR and EMLLR adapted samples from speakers M1, M2, F1 and F3.

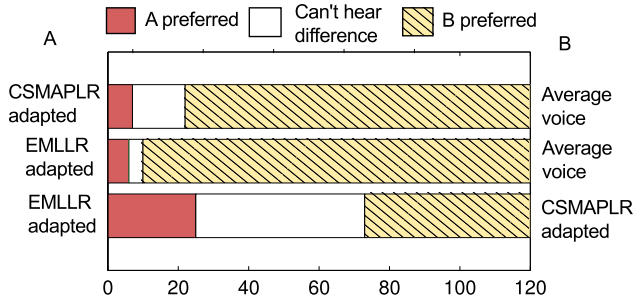
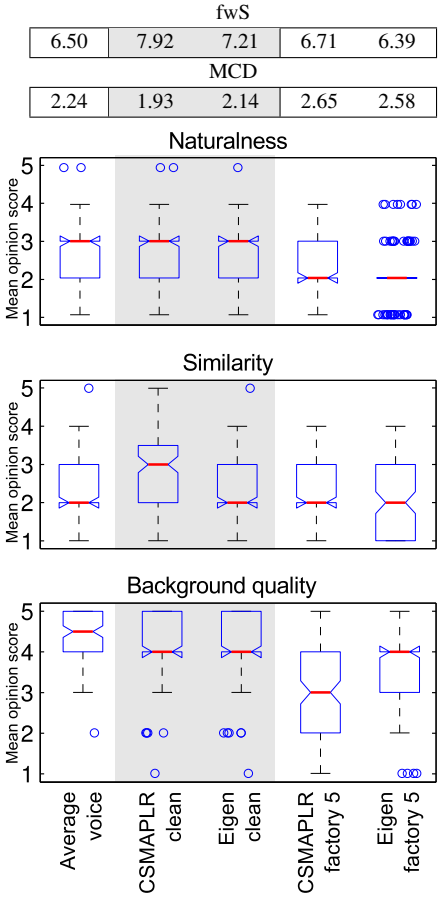


Fig. 7. Results of the listening test B task 2 with natural and synthetic samples from speaker M1, M2, F1 and F3. Box plots described in Figure 3.



average voices were also included in the test.

Results from the AB test where listeners compared samples generated with the average voice models to samples generated with models adapted using noisy data are presented in Figure 6. The adaptation data was corrupted with factory noise at SNR 5 dB. The results clearly indicate that listeners prefer a good quality average voice to represent the reference speaker in real world application ($p < 0.001$). Additionally, samples generated with CSMAPLR-adapted models are considered better than samples generated with EMLLR-adapted models, but the difference is not conclusive ($p = 0.21$).

The MOS evaluation results (Figure 7) indicate that when noisy adaptation data is used, the eigen-adapted models produce cleaner speech than models adapted with CSMAPLR. The speaker similarity scores are better with CSMAPLR, but when noisy adaptation data is used, the difference between the adaptation methods is small and neither adaptation method can improve speaker similarity compared to the average voice model. The perceived naturalness appears to depend on whether the adaptation data was clean or noisy, whereas in the other test, the naturalness scores were almost same for all synthesised samples (Figure 3).

The objective evaluation measures calculated for the listening test samples are in agreement with subjective evaluations. Adaptation with clean data results in better scores than adaptation with noisy data, and when clean adaptation data is used, CSMAPLR samples receive the best scores. When noisy adaptation data is used, fwS scores are best with CSMAPLR adaptation whereas MCD scores are better with EMLLR adaptation. MCD scores are better for samples generated with the average voice model than for samples generated with the adapted models when noisy adaptation data is used, which corresponds to the subjective AB test results. The linear correlation coefficients between the average fwS and MOS scores in naturalness, speaker similarity, and background quality are $|\rho| = 0.63$, $|\rho| = 0.83$, and $|\rho| = 0.50$, and between the MCD and MOS scores $|\rho| = 0.93$, $|\rho| = 0.86$, and $|\rho| = 0.90$.

VII. Conclusions and Discussion

In this work, we have described our experiments on using noisy data in HMM-TTS adaptation [6] in greater detail and examined how noise affects the different vocoder features. The results discussed in Section V-B suggest that in STRAIGHT vocoding, the excitation parameters are less prone to degradation by noise than the cepstral filter parameters. The objective and subjective evaluation additionally indicate that using NMF-based speech enhancement on the noisy adaptation data introduces some improvement in the synthesised speech quality, but the difference between systems adapted from noisy and enhanced data in the AB test was not statistically significant.

As discussed in [6], the baseline speaker-adaptive HMM-TTS system with CSMAPLR/CMLLR adaptation is relatively robust to environmental noise when enough data is available. In this work, we further studied the relation between adaptation data size and synthesised speech quality. The results suggested that when the data is corrupted with occasional noise bursts, using more adaptation data can compensate for the degradation due to noise, but when noise is continuous, using more data data does not result in a notable improvement in quality. That is, a small set of clean adaptation data will result in a better quality synthesised speech than a large set of noisy data if the noise is continuous.

To improve adaptation performance in noisy conditions, we tested using a noise-robust feature stream for model alignment but marked that adaptation of the standard STRAIGHT-based acoustic features yields equally usable state alignments. We also attempted noise-dampening adaptation with EMLLR

adaptation but discovered that even though this adaptation technique is more robust towards noise than standard MLLR-based adaptation mechanisms, the quality reached is not yet satisfactory. The subjective listening test results reported in Section VI-D indicate that for real-world applications, representing the target speaker with a generic male or female voice is preferred over CSMAPLR or EMLLR adaptation from a small amount of noisy utterances.

Evaluation methods proposed in [6] were used throughout this work, and correlation between the subjective and objective evaluations was analysed. The results suggest the background quality effects the frequency-weighted segmental SNR less than mel-cepstral distortion, and in the second listening test, the objective measures rank CSMAPLR and EMLLR adaptated models in different order when noisy adaptation data is used. We note that the frequency-weighted segmental SNR can also be calculated based on STRAIGHT rather than FFT spectra, and this can introduce a modest improvement to the correlation with subjective evaluation results, as reported in [39].

In conclusion, although the features and adaptation methods used in HMM-TTS systems exhibit a level of noise robustness, using noisy adaptation data degrades the synthesised speech quality in a manner that cannot be compensated for with more data. With the evaluation tools and baseline results presented in this work, we can turn our gaze into a variety of technological questions. In the neighbouring fields of noise-robust ASR and noise-removal in telecommunication applications, a variety of techniques have been developed to improve noise robustness. While many may not be applicable to HMM-synthesis due to the requirements for the personal and natural qualities of the voice are high, techniques such as non-linear kernel methods for adaptation and missing data methods for feature extraction provide possibilities for future work. Finally, we note that given the option to use a large amount of noisy data or discard noisy utterances and use a small amount of clean data, the choice does not need to be exclusive. For example, all data could be used for F0 extraction and adaptation, whereas only selected high-quality utterances would be used to adapt the MCEP parameters.

ACKNOWLEDGEMENTS

We acknowledge the computational resources provided by Aalto Science-IT project.

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 Trans. Audio, Speech, Lang. Process*, 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. Information and Systems*, 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] J. Yamagishi, Z. Ling, and S. King, "Robustness of HMM-based speech synthesis," in *Proc. Interspeech*, 2008.

[6] R. Karhila, U. Remes, and M. Kurimo, "HMM-based speech synthesis adaptation using noisy data: Analysis and evaluation methods," in *Proc. ICASSP*, 2013.

[7] 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.

[8] K.-T. Chen, W.-W. Liau, H.-M. Wang, and L.-S. Lee, "Fast speaker adaptation using eigenspace-based maximum likelihood linear regression," in *Proc. ICSLP*, 2000.

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

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

[11] 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.

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

[13] R. Karhila and M. Wester, "Rapid adaptation of foreign-accented HMM-based speech synthesis," in *Proc. Interspeech*, 2011.

[14] R. Kuhn, J.-C. Junqua, P. Nguyen, and N. Niedzielski, "Rapid speaker adaptation in eigenvoice space," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 6, pp. 695–707, 2000.

[15] K. Shichiri, A. Sawabe, T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Eigenvoices for HMM-based speech synthesis," in *Proc. ICSLP*, vol. 2, 2002, pp. 1269–1272.

[16] S. Mika, B. Schölkopf, A. Smola, K.-R. Müller, M. Scholz, and G. Rätsch, "Kernel PCA and de-noising in feature spaces," in *Proc. NIPS*, 1999, pp. 536–542.

[17] M. J. F. Gales, "Cluster adaptive training of hidden Markov models," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 4, pp. 417–428, 2000.

[18] H. Zen, N. Braunschweiler, S. Buchholz, M. J. F. Gales, K. Knill, S. Krstulovic, and J. Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 20, no. 6, pp. 1713–1724, 2012.

[19] V. Wan, J. Latorre, K. Chin, L. Chen, M. Gales, H. Zen, K. Knill, and M. Akamine, "Combining multiple high quality corpora for improving HMM-TTS," in *Proc. Interspeech*, 2012.

[20] B. Mak, J. Kwok, and S. Ho, "Kernel eigenvoice speaker adaptation," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5, pp. 984–992, 2005.

[21] B. K.-W. Mak, R. W.-H. Hsiao, S. K.-L. Ho, and J. Kwok, "Embedded kernel eigenvoice speaker adaptation and its implication to reference speaker weighting," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 14, no. 4, pp. 1267–1280, july 2006.

[22] Z. Roupakia and M. Gales, "Kernel eigenvoices (revisited) for large-vocabulary speech recognition," *IEEE Signal Processing Letters*, vol. 18, no. 12, pp. 709–712, dec. 2011.

[23] B. Mak and R. Hsiao, "Kernel eigenspace-based MLLR adaptation," *IEEE Trans. Audio, Speech, Lang. Process*, pp. 784–795, 2007.

[24] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE J. Selected Areas in Communications*, vol. 6, no. 2, pp. 242–248, 1988.

[25] J. M. Tribolet, P. Noll, B. J. McDermott, and R. E. Crochiere, "A study of complexity and quality of speech waveform coders," in *Proc. ICASSP*, 1978, pp. 586–590.

[26] 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*.

[27] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 16, no. 1, pp. 229–238, jan. 2008.

[28] M. Brookes, "VOICEBOX: Speech processing toolbox for MATLAB," 1998.

[29] R. A. J. Clark, M. Podsiadło, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proc. Blizzard Workshop*, 2007.

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

[31] A. Varga and H. 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.
[32] J. Yamagishi and O. Watts, "The CSTR/EMIME HTS system for Blizzard Challenge," in *Proc. Blizzard Challenge*, 2010.
[33] H. Kawahara, I. Masuda-Katsuse, and A. de 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.
[34] D. Arifianto, T. Tanaka, T. Masuko, and T. Kobayashi, "Robust F0 estimation of speech signal using harmonicity measure based on instantaneous frequency," *IEICE Trans. Inf. Syst.*, vol. 87, no. 12, pp. 2812–2820, 2004.
[35] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. Eurospeech*, 1999, pp. 2781–2784.
[36] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, pp. 495–518, 1995.
[37] ETSI, *ES 202 050 V1.1.5 Speech processing, transmission and quality aspects (STQ), distributed speech recognition*, 2007.
[38] D. Iskra, B. Grosskopf, K. Marasek, H. van den Heuvel, F. Diehl, and A. Kiessling, "Speecon-speech databases for consumer devices: Database specification and validation," in *Proc. LREC*, 2002.
[39] U. Remes, R. Karhila, and M. Kurimo, "Objective evaluation measures for speaker-adaptive HMM-TTS systems," in *Proc. SSW*, to appear.

Reima Karhila (M'11) received the M.Sc. degree in communications engineering from the Aalto University (former Helsinki University of Technology), Finland, in 2010 He is currently pursuing the D.Sc. (Tech.) degree at the Department of Signal Processing and Acoustics at Aalto University.

His current research interest is acoustic modelling in statistical parametric speech synthesis using substandard data.

Ulpu Remes received the M.Sc. degree in engineering physics from the Helsinki University of Technology, Espoo, Finland, in 2007. She is currently pursuing the D.Sc. (Tech.) degree at the Department of Signal Processing and Acoustics at Aalto University.

Her research interests include machine learning and noise-robust speech recognition.

Mikko Kurimo (SM'07) received the Dr.Sc. (Ph.D.) in technology degree in computer science from the Helsinki University of Technology, Espoo, Finland, in 1997. He is currently Associate Professor in the Department of Signal Processing and Acoustics at Aalto University, Finland.

His research interests are in speech recognition and synthesis, machine learning, information retrieval, natural language processing, and multimodal interfaces.