Objective evaluation measures for speaker-adaptive HMM-TTS systems

Ulpu Remes, Reima Karhila, Mikko Kurimo

Department of Signal Processing and Acoustics, Aalto University School of Electrical Engineering, Finland

firstname.lastname@aalto.fi

Abstract

This paper investigates using objective quality measures to evaluate speaker adaptation performance in HMM-based speech synthesis. We compare several objective measures to subjective evaluation results from our earlier work about 1) comparison of speaker adaptation methods for child voices and 2) effects of noise in speaker adaptation. The results analysed in this work indicate a reasonable correlation between several objective and subjective quality measures.

Index Terms: adaptation, speech synthesis, evaluation

1. Introduction

Hidden Markov model (HMM) based text-to-speech (TTS) framework [1] is an attractive alternative to conventional concatenative speech synthesis. While concatenative systems typically produce natural and understandable speech, HMM-TTS systems are more flexible and can be adapted to mimic different speaking styles or speakers based on a limited amount of adaptation data [2].

Speaker adaptation performance in HMM-TTS systems is typically evaluated using subjective listening tests [3]. The samples generated with speaker-adapted models are rated based on whether the synthesised voice sounds like the target speaker and on the perceived naturalness. While listening tests are necessary to confirm how differences between adaptation methods are perceived and appreciated, subjective evaluation is not an efficient tool for tasks such as parameter tuning that require iterative evaluation.

Numerous objective quality measures have been developed for speech quality evaluation in telecommunication systems [4, 5]. In addition to telecommunication systems, the measures have been used to evaluate speech enhancement systems and have correlated well with subjective evaluations [6]. While the degradation introduced in speech transmission or enhancement may have a fundamentally different nature compared to samples generated with statistical speech synthesis [7, 8], the same objective measures have been applied to evaluate speech synthesis systems [9, 8, 10, 11].

In this work, we replicate and expand the previous studies [6, 10, 11] on correlation between objective quality measures and subjective listening test results. We focus on subjective evaluations of speaker-adaptation performance in HMM-TTS systems, using limited sets of data from our earlier works [12, 11]. Beside comparing objective measures, we investigate, for the spectrum-based measures, whether measuring the spectrum as synthesised or analysed after a complete waveform synthesis affects the measures.

The rest of the paper is organised as follows. In Section 2, we describe the objective evaluation measures used in this work. In Section 3, we revisit the listening test results used in previous research and evaluate the correlation between objective and subjective evaluations. All the speech synthesis systems represented in the results are personalised HMM-TTS systems. The results are discussed in Section 4, and Section 5 concludes the work.

2. Methods

2.1. Objective measures

HMM-based speech synthesis systems use statistical models to generate the spectral envelope and F_0 contour that are input to a vocoder that synthesises the final output waveform. To evaluate the system performance, the estimated parameters can be compared with equivalent parameters extracted from a reference speech sample [13, 2]. The mel-cepstral distance (MCD) is calculated as

$$MCD = \frac{1}{M} \sum_m \sqrt{2 \sum_d (c(d,m) - \hat{c}(d,m))^2} \qquad (1)$$

where $\hat{c}(d,m)$ and $c(d,m)$ denote the dth mel-cepstral coefficient of the test and reference signals in time frame m and M denotes the number of frames. We note that a synthesised test sample can be represented with the internal mel-cepstrum which is used as a synthesis parameter or with mel-cepstral coefficients extracted from the synthesised output. We compute and evaluate both internal and output mel-cepstral distances.

The other evaluation measures used in this work have been developed to assess the speech enhancement or transmission qualities. We focus on the frequency-weighted segmental SNR (FWS) [4] that exhibited a performance close standardised PESQ objective evaluation measure in a speech enhancement evaluation task [6]. We calculate the frequency-weighted segmental SNR in mel-spectral domain as

$$FWS = \frac{1}{M} \sum_m \sum_k W(k,m) \log_{10} \frac{X(k,m)^2}{(X(k,m) - \hat{X}(k,m))^2} \qquad (2)$$

where $\hat{X}(k,m)$ and $X(k,m)$ denote the kth mel-spectral component of the test and reference samples in time frame m. As proposed in [6], the mel-spectrum in each time frame m is normalised to unit area ($\sum_k X(k,m) = 1$) and the channels are weighted as

$$W(k,m) = X(k,m)^\gamma / \sum_k X(k,m)^\gamma, \qquad (3)$$

where $\gamma = 0.2$. The mel-spectral features $\hat{X}(k,m)$ that represent a synthesised test sample are calculated based on the internal mel-cepstral representation or extracted from the synthe-

sised output as discussed in Section 2.2. The estimated signal-to-noise ratio in each time frame is bound to $[0, 35]$ dB range as proposed in [14].

We additionally calculate the cepstral distance (CEP), log-likelihood ratio (LLR) [15], and weighted spectral slope (WSS) measure [16] using the implementations in COLEA toolbox [17]. The measures calculated with COLEA are computed based on the reference sample and synthesised output samples in time-domain. LLR is calculated based on order 10 linear prediction models.

2.2. Feature extraction

To compare the reference samples to samples generated with the HMM-TTS system, the synthesised samples were generated based on the phone alignment of the reference sample. The synthesised samples were associated with the internal mel-cepstral and spectral representation that correspond to the mel-cepstral and spectral features generated with the STRAIGHT vocoder [18]. MCD and FWS calculated based on the internal representations and STRAIGHT-based representations generated from the synthesised output are compared in this work. We additionally calculate the FWS measure based on FFT spectra calculated for the test and reference samples to compare the FFT and STRAIGHT spectrum.

The test and reference samples have 16 kHz sampling rate. The samples were processed in 25 ms Hamming windows with 5 ms shift between adjacent frames. The mel-filterbank applied to FFT or STRAIGHT spectra was estimated with VOICEBOX [19].

The objective measures were calculated based on 2 second samples extracted from the middle of the utterance as proposed in [11]. For some reason, the synthesis system used in this work occasionally introduces excess frames in the beginning or end of the sample. The comparison between the reference and test samples was therefore done at several frame delays $[-10 \ldots 10]$ and the best match was recorded.

3. Evaluation

3.1. Subjective test data

The objective measures are compared with mean opinion scores (MOS) collected in two speaker-adaptive HMM-TTS evaluations [12, 11]. Finnish speech data was used in both evaluations. The synthesis systems used in one evaluation share a common framework: the same prosody-prediction and model selection front-end are used and the phoneme sets are identical. The system parameters and training are described in the original papers [12, 11].

The first evaluation compared speaker adaptation methods for child voices. The differences between the synthesis systems in this evaluation stem from differences in the average voice training database and adaptation procedures [12]. Adaptation performance was evaluated in a subjective listening test where 26 listeners rated test samples from three target speakers based on similarity with the target speaker and naturalness. The samples were evaluated on a scale of 1–5. The mean opinion scores are reported in Figure 1.

The second evaluation studied using noisy and enhanced speech data for speaker adaptation in HMM-TTS systems. The differences between the synthesis systems stem from the differences in the adaptation training data that was either clean, noise-corrupted, or enhanced [11]. The evaluation focussed on the mel-cepstrum and excitation components, which were generated with the HMM-TTS system, whereas the F0 contours for synthesis were extracted from the reference samples. Adaptation to one female and one male speaker was evaluated, and the listening test samples included also noise-corrupted and enhanced natural speech samples. 26 listeners evaluated the samples based on their naturalness, similarity, and background intrusiveness as proposed in [11]. The subjective evaluation scales are described Table 1 and the mean opinion scores reported in Figure 2.

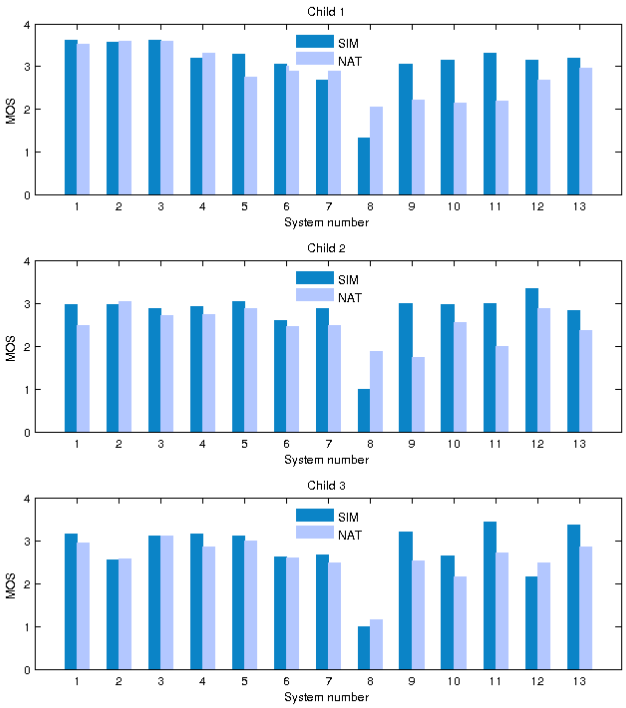

Figure 1: Mean opinion scores (MOS) on synthesised child speech. Thirteen systems were evaluated for similarity (SIM) and naturalness (NAT).

3.2. Evaluation measures

Concordance between the subjective and objective measures is assessed with a sample correlation coefficient $|\bar{r}|$. The standard sample correlation r is modified to marginalise the level differences between scores assigned to individual speakers n and emphasise the comparison between the tested conditions or systems. The modified sample correlation coefficient is calculated as

$$\bar{r} = \frac{1}{N} \sum_{n} r(n) \tag{4}$$

where N denotes the number of speakers and $r(n)$ is the speaker-conditioned sample correlation. The sample correlation between the subjective and objective scores assigned to speaker n is calculated as

$$r(n) = \frac{\sum_i (S_i(n) - \bar{S}(n))(O_i(n) - \bar{O}(n))}{\sqrt{\sum_i (S_i(n) - \bar{S}(n))^2 \sum_i (O_i(n) - \bar{O}(n))^2}} \tag{5}$$

Table 1: Subjective listening test scales

	Similarity (SIM)
5	Exactly like the same person
4	Quite like the same person
3	Somewhat different but recognisable as the same person
2	Quite like a different person
1	Like a totally different person

	Naturalness (NAT)
5	Completely natural
4	Quite natural
3	Somewhat unnatural but acceptable
2	Quite unnatural
1	Completely unnatural

	Background (BAK)
5	Clean
4	Quite clean
3	Somewhat noisy but not intrusive
2	Quite noisy and somewhat intrusive
1	Very noisy and very intrusive

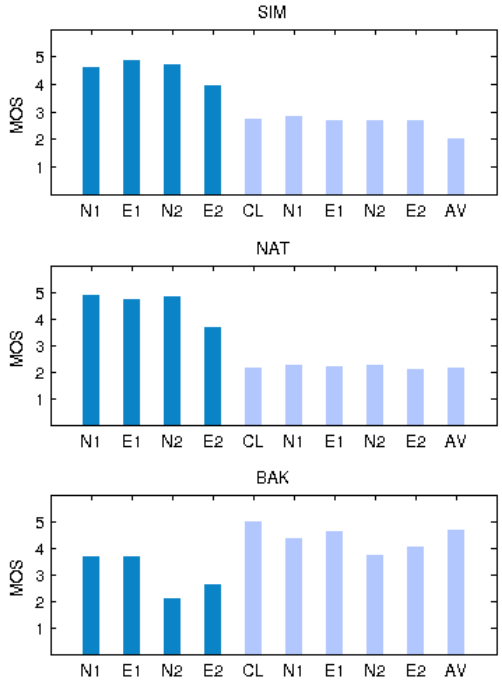

Figure 2: Mean opinion scores (MOS) on natural speech samples (dark colour) that have been corrupted with noise (N1–N2) and enhanced (E1–E2) and synthesised samples generated with HMM-based TTS. The synthesised samples represent the average voice model (AV) and models adapted with clean (CL) and noise-corrupted and enhanced data.

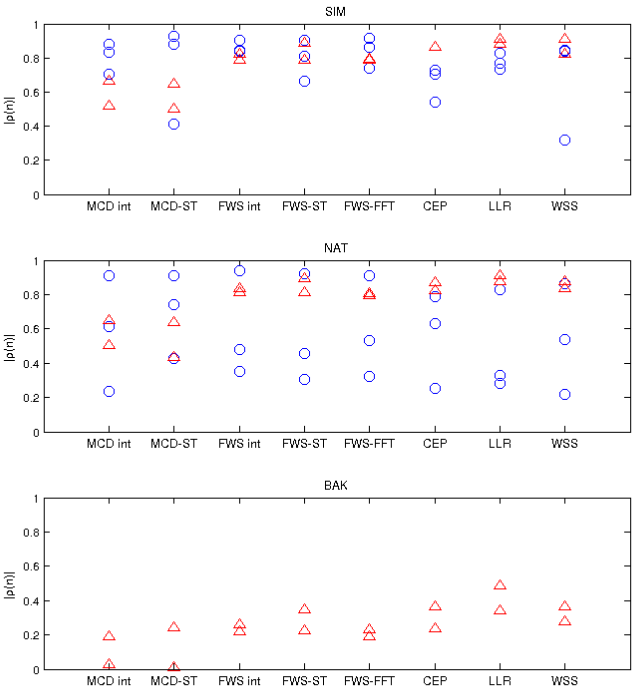

Figure 3: Linear correlation coefficients $|\rho(n)|$ between the objective evaluation measures and the subjective mean opinion scores. Results related to the first test with three child voices are indicated with circles and results related to the second test with one male and one female voice with are indicated with triangles.

where $S_i(n)$ and $O_i(n)$ denote the subjective and objective score calculated for test sample (i, n) and $\bar{S}(n)$ and $\bar{O}(n)$ the mean of the subjective and objective scores assigned to speaker n, $\bar{S}(n) = \sum_i S_i(n)$ and $\bar{O}(n) = \sum_i O_i(n)$. The subjective scores $S_i(n)$ are average scores calculated across the individual listener ratings for test sample (i, n). Listener ratings that deviated more than two standard deviations from the test sample mean were discarded as outliers and are not reflected in the averages $S_i(n)$.

3.3. Results

The sample correlations $|r(n)|$ between the objective measures and subjective listening test results are reported in Figure 3. We have evaluated the correlation between subjective evaluations and (a) MCD calculated based on the internal STRAIGHT representation and (b) STRAIGHT representation extracted from the synthesised output, (c) FWS calculated based on the internal representation, (d) representation extracted from the synthesised output and (e) FFT spectrum calculated from the synthesised output, (f) cepstral distance, (g) LLR and (h) WSS. FWS measures were calculated based on $K = 13$ and $K = 21$ mel-channels, but the differences in the results and in their correlation with MOS scores were small. The results reported in Figure 3 pertain to FWS measures calculated based on $K = 21$ mel-channels.

The sample correlations between the objective measures and subjective listening test results on the synthesised child voices are indicated with circles in Figure 3. We note that the subjective SIM and NAT evaluations are inter-dependent to certain extent, but their relationship is not linear ($|\bar{r}| = 0.65$). The objective measures evaluated in this work correlate better with the similarity evaluations. The best sample correlation with SIM ($|\bar{r}| = 0.86$) is obtained with the FWS measure calculated based on the internal STRAIGHT representation and the best sample correlation with NAT ($|\bar{r}| = 0.69$) with MCD calculated based on the synthesised output.

When noise-corrupted or enhanced samples or synthesis models adapted with noise-corrupted or enhanced data were used, the samples were evaluated in three scales (Table 1). The sample correlations between the objective measures and subjective evaluations are indicated with triangles in Figure 3. The notable quality difference between the natural and synthesised samples dominates the SIM and NAT evaluations which are exceptionally coherent across the test conditions ($|\bar{r}| = 0.97$). The best correlation with SIM ($|\bar{r}| = 0.90$), NAT ($|\bar{r}| = 0.89$), and BAK ($|\bar{r}| = 0.41$) is obtained with LLR.

While correlation between the objective evaluations and BAK appears weak when examined over the complete test set, background intrusiveness has a notable contribution to the objective scores. The sample correlation calculated for objective measures and BAK $|r(n)| \geq 0.93$ within the natural sample set and $|r(n)| \geq 0.70$ within the synthesised samples that represent an adapted model. This suggests that the objective measures emphasise speech qualities but are not invariant to background intrusiveness.

4. Discussion

4.1. Main results

We evaluated the correlation between several objective measures and subjective listening test results in two tasks. FWS calculated based on the internal spectral representation and LLR resulted in the best overall correlation with the subjective similarity scores. As discussed in [8], the correlation between the objective measures and subjective evaluations varies from voice to voice, but the sample correlation calculated between SIM and FWS int for individual voices in either dataset $|r(n)| > 0.75$ and the sample correlation between SIM and LLR measures $|r(n)| > 0.73$. FWS calculated based on the synthesised output also correlated well with SIM evaluations, and we note that FWS and LLR performed well also in the speech enhancement evaluation [6].

In the previous studies [9, 6, 10], the best correlation with subjective listening test results has been obtained with PESQ [5]. This is a standardised measure that incorporates perceptual and cognitive models for speech quality assessment. Despite the success obtained with PESQ, we believe the need for license-free evaluation measures remains. With projects like Simple4All [1], HMM-TTS is becoming more accessible for languages that are under-resourced both in terms of data an funding.

4.2. Similarity and naturalness

An ideal measure for objective evaluation would take into account all the factors that influence subjective listening test results, but this is a difficult task. For example, a smooth and natural voice is often rated better than a rougher voice that is otherwise more similar to the target speaker. Evaluation in a noisy background adds to the complexity as the increased background noise can mask synthesis artifacts and make a synthesised voice sound better.

The objective quality measures evaluated in this work operate on a one-dimensional scale whereas the human listeners rated the samples on based on two or three specific features. This is necessary when several factors affect the perceived overall quality. Hu and Loizou [6] used linear combinations of the basic objective measures to calculate composite measures tuned for the separate subjective scales. The measures evaluated in this work are, however, very correlated, which means nonlinear combinations should be used in order to introduce notable improvement compared to the best individual measures. The similarity aspect could also be assessed with speaker recognition techniques, for example.

4.3. Reference data

The objective measures evaluated in this work represent the so called intrusive or full-reference measures that require a target sample for comparison. Therefore the synthesised samples had to be generated with the alignments extracted from the target signal. Möller et al. [8] compared three non-intrusive measures in speech transmission and speech synthesis evaluation task, but concluded that the objective measures were not sufficiently accurate in predicting differences between synthesised speech quality.

Developing model-based measures for speaker similarity and naturalness would allow us to evaluate the quality of synthetic speech with less than perfect time alignment between the reference and synthetic stimulus. This will be growingly more desirable as synthetic speech tries to reproduce the prosodic aspects of the speech, including accent and speaking rhythm. Evaluating prosody with objective measures is not a realistic goal in the near future, but as the development prosody generation and spectral envelope synthesis can not be done completely separately, also the objective evaluation of the spectral envelope should be done in conjunction with model-generated prosody.

5. Conclusions and future work

We analysed correlation between objective measures and subjective listening test results in speaker adaptation task in the HMM-TTS framework. The measures were correlated with the subjective speaker similarity more than naturalness or background quality, and the best measures were FWS and LLR. While the measures studied in this work cannot replace subjective evaluation, the measures could be used to optimise the system parameters in a manner that better corresponds to listener preference. For example, speech enhancement parameters and regression tree size in adaptation are usually hand-tuned based on performance over some development data, and any measure can be used for the performance evaluation. MCD can also be used to optimise the adaptation transformations as proposed in [20].

Our datasets were quite small, and we would like to continue the verification process using larger datasets. We think objective evaluation should not require aligned samples, and therefore hope to investigate if test and reference samples could be compared based on local alignments, and if speaker recognition technologies could provide measures that correlate with the similarity scores in our listening tests.

[1] http://www.simple4all.org

6. Acknowledgements

This work received financial support from the Academy of Finland under the grants no 135003, 140969, and 251170, from Tekes under Perso and Funesomo projects, and from EC FP7 under grant agreement 287678. R. Karhila was supported by Langnet graduate school and Nokia Foundation. 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 Trans. Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 984–1004, 2010.

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

[4] 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.

[5] 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*.

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

[7] N. Kitawaki and H. Nagabuchi, "Quality assessment of speech coding and speech synthesis systems," *IEEE Communications Magazine*, vol. 26, pp. 36–44, 1988.

[8] S. Möller, D. S. Kim, and L. Malfait, "Estimating the quality of synthesized and natural speech transmitted through telephone networks using single-ended prediction models," *Acta Acustica united with Acustica*, vol. 94, pp. 21–31, 2008.

[9] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the pesq measure," in *Proc. European Congress on Acoustics*, 2005, pp. 2725–2728.

[10] D. Y. Huang, "Prediction of perceived sound quality of synthetic speech," in *Proc. APSIPA*, 2011.

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

[12] R. Karhila, R. S. Doddipatla, M. Kurimo, and P. Smit, "Creating synthetic voices for children by adapting adult average voice using stacked transformations and VTLN," in *Proc. ICASSP*, 2012.

[13] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, Lang. Process*, vol. 17, no. 1, pp. 66–83, 2009.

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

[15] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 23, pp. 67–72, 1975.

[16] D. H. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: a first step," in *Proc. ICASSP*, 1982, pp. 1278–1281.

[17] P. Loizou, "COLEA: A MATLAB tool for speech analysis," 1998.

[18] 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.

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

[20] L. Qin, Y. J. Wu, Z. H. Ling, R. H. Wang, and L. R. Dai, "Minimum generation error linear regression based model adaptation for HMM-based speech synthesis," in *Proc. ICASSP*, 2008, pp. 3953–3956.