ASSESSMENT AND CORRECTION OF VOICE QUALITY VARIABILITIES IN LARGE SPEECH DATABASES FOR CONCATENATIVE SPEECH SYNTHESIS

Yannis Stylianou

AT&T Labs-Research, SIPS, 180 Park Avenue, Florham Park, NJ 07932
email : styliano@research.att.com

ABSTRACT

In an effort to increase the naturalness of concatenative speech synthesis, large speech databases may be recorded. While it is desirable to have varied prosodic and spectral characteristics in the database, it is not desirable to have variable voice quality. In this paper we present an automatic method for voice quality assessment and correction, whenever necessary, of large speech databases for concatenative speech synthesis. The proposed method is based on the use of a Gaussian Mixture Model, GMM, to model the acoustic space of the speaker of the database and on autoregressive filters for compensation. An objective method to measure the effectiveness of the database correction based on a likelihood function for the speaker's GMM, is presented as well. Both objective and subjective results show that the proposed method succeeds in detecting voice quality problems and successfully corrects them. Results show a 14.2% improvement of the log-likelihood function after compensation.

1. INTRODUCTION

While good quality speech synthesis is obtained using concatenation of a small set of controlled units (e.g. diphones) the availability of more units taken from large speech databases seems to be the key for natural sounding text-to-speech systems. Increasing the size of the database one increases the instances of the basic units (phonemes, diphones, etc.) that a synthesizer can use. Then, to synthesize a text sentence, an approach known as unit selection[1, 2], is needed to find an optimum set of basic units that matches the desired prosodic features of the sentence. Thus the variety of prosodic characteristics and spectral variations of the same type of basic unit will either reduce the prosodic modifications that a signal processing module needs to perform or eliminate any prosodic modification (in case that the selected unit has prosodic and spectral characteristics very close to the desired ones). By removing the necessity of extended prosodic modifications, a higher naturalness of the synthetic speech is achieved. While having many different tokens for each basic unit is strongly desired, a variable voice quality is not; if it exists, it will not only make the concatenation task more difficult but also will result in a synthetic speech with changing voice quality even within the same sentence. Depending on the variability of the voice quality of the database a synthetic sentence can be perceived from being "rough" (even if a smoothing algorithm is used at each concatenation instant), and, in the worst case, as if different speakers utter various parts of a sentence. Thus, inconsistencies in voice quality within the same unit-selection speech database can degrade the overall quality of the synthesis. On the other hand, if the unit selection procedure is highly discriminative, it will exclude for concatenation units of the database with a mismatch in voice quality. Then, however, the synthesizer will only use part of the database while time (and money) was invested to make the complete database available (recording, phonetic labeling, prosodic labeling, etc.). In this case, increasing the size of the database will not necessarily increase proportionally the instances of units available for synthesis.

Recording large speech databases for speech synthesis is a very long process with duration from many days to months. The duration of each recording session can be as long as 5 hours (including breaks, instructions, etc.) and the time between contiguous recording sessions can be more than a week. Thus, the probability to have variations in voice quality from one recording session to another (*inter-session variability*) as well as during the same recording session (*intra-session variability*) is high. A reason for the inter-session variability could be associated with a different emotional or health situation of the speaker (e.g., an upcoming cold) or with a difference in the recording equipment that is used (microphone placement, amplifier settings, etc.) The main reason for the intra-session variability is the fatigue of the speaker because of the length of the recording session.

The problem of voice quality assessment for a given speech database seems to have similarities with the speaker adaptation problems in speech recognition. There, "data oriented" compensation techniques have been proposed that attempt to filter noisy speech feature vectors to produce "clean" speech feature vectors [3]. However, in recognition it is the recognition score that is of interest regardless of whether or not the adapted speech feature vector really matches that of "clean" speech. Another domain which has similarities with our problem is speech enhancement. Although in this case the output is speech, the quality of the enhanced speech is not as high as truly clean speech. The above discussion clearly shows the difficulty of our problem: not only an automatic detection of quality is sought (so if there is no need for correction the speech signal will not be modified) but in addition any modification/correction of the signal has to result in speech of (extremely) high quality. Otherwise the overall attempt to correct the database has no meaning for speech synthesis. While consistency of voice quality in a unit-selection speech database is, therefore, important for high-quality speech synthesis, no method for automatic voice quality assessment and correction in the context of text-to-speech synthesis has been proposed yet.

In this paper we propose an automatic method for detection of voice quality problems in a large speech database, as well as a correction method that preserves high quality. The detection method is based on probabilistic criteria and makes use of a Gaussian Mixture Model (GMM) for modeling the acoustic space of the speaker of the database. The part of the database used for modeling is as-

sumed to be of high and *accepted* speech quality. The meaning of the term *accepted* within our application will be clarified in the next sections. Building on the accepted foundation, the likelihood that the estimated GMM has generated other parts of the database is estimated and compared with a lower and an upper bound. For each of the areas with low likelihood a corrective filter is estimated based on linear prediction and average periodograms. Likelihood values found before and after corrections are very correlated with subjective results.

2. DETECTION OF VOICE QUALITY VARIABILITY

In this section we consider the detection of voice quality variability. To represent the speaker of the database a Gaussian Mixture Model (GMM) is used. The GMM is a parametric model successfully applied in speaker identification [4]. A Gaussian mixture density assumes that the probability distribution of the observed parameters, $\mathbf{O}$, is given by the following equation,

$$p(\mathbf{O}|\Lambda) = \sum_{i=1}^{M} \alpha_i p(\mathbf{O}|\lambda_i), \tag{1}$$

where M is the number of the Gaussian components, α_i represents the statistical frequency of each class in the observations and $p(\mathbf{O}|\lambda_i)$ denotes the p-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The complete Gaussian mixture density is represented by the model,

$$\Lambda = \{\lambda_i\} = \{\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \text{ for } i = 1, \ldots, M. \tag{2}$$

Let r_n, $n = 1, \ldots, N$ denote different recording sessions and r_p be the recording session with the preferred voice quality (reference recording session). We denote the set of L observation sequences from r_p as,

$$\mathbf{O}_{r_p} = \left[\mathbf{O}_{r_p}^{(1)}, \mathbf{O}_{r_p}^{(2)}, \ldots, \mathbf{O}_{r_p}^{(k)} \vdots \mathbf{O}_{r_p}^{(k+1)}, \ldots, \mathbf{O}_{r_p}^{(L)}\right] \tag{3}$$

where $\mathbf{O}_{r_p}^{(l)} = (\mathbf{o}_1^{(l)} \mathbf{o}_2^{(l)} \ldots \mathbf{o}_T^{(l)})$ is the lth observation sequence. The parameters of GMM, Λ_{r_p}, are estimated from the first k observations $\left[\mathbf{O}_{r_p}^{(1)}, \mathbf{O}_{r_p}^{(2)}, \ldots, \mathbf{O}_{r_p}^{(k)}\right]$ using the Expectation-Maximization (EM) algorithm [5]. Given the model Λ_{r_p} the *log likelihood functions*

$$\mathcal{L}(\mathbf{O}_{r_p}^{(l)}|\Lambda_{r_p}) = \frac{1}{T}\sum_{t=1}^{T} p(\mathbf{o}_t^{(l)}|\Lambda_{r_p}) \tag{4}$$

are estimated for $l = 1, \ldots, L$, where $p(\mathbf{o}_t^{(l)}|\Lambda_{r_p})$ is given from (1).

The log likelihood function is a measure of how likely it is that the model Λ_{r_p} has produced the set of observed samples. Using as learning set the first k observations and the *whole* reference recording session, r_p, as test data, upper and lower bounds for the log likelihood function, $\mathcal{L}$, can be obtained. The distribution of $\mathcal{L}$ for the entire r_p can be approximated with a uni-modal Gaussian with mean $\mu_{\mathcal{L}}$ and variance $\sigma_{\mathcal{L}}^2$.

Voice quality problems of the other recording sessions are then detected by computing the z-score[1] of the log likelihood function of observations from these sessions regarding the model Λ_{r_p},

$$z_{r_i}^{l} = \frac{\mathcal{L}(\mathbf{O}_{r_i}^{(l)}|\Lambda_{r_p}) - \mu_{\mathcal{L}}}{\sigma_{\mathcal{L}}} \tag{5}$$

where $\mathbf{O}_{r_i}^{(l)}$ denotes the lth observation from the r_i, $(i \neq p)$ recording session. Thus, detection turns into a hypothesis testing problem. The two hypotheses are the *null* hypothesis, denoted by H_0: r_p and the lth observation from r_i have the same voice quality ($r_p \sim r_i(l)$), and the *alternative* hypothesis denoted by H_1: r_p and the lth observation from r_i have different voice qualities ($r_p \not\sim r_i(l)$). The level of alpha error (the probability of being wrong whenever the null hypothesis is rejected, or Type 1 error) used was 0.01.

3. COMPENSATION

For each part of the database where the hypothesis H_0 has been rejected a corrective filter is assigned. While the characteristics of unvoiced speech differ from those of voiced speech, it was decided to use the same correction filter for both cases. This is motivated by the fact that the system tries to detect and correct *average* differences in voice quality. For a subset of our applications (e.g., detection of different microphone positions), this kind of variability is assumed to be identical for voiced and unvoiced sounds. In other cases, for example, if we aim at detecting speaker fatigue at the end of a recording session, voiced and unvoiced sounds might be affected in different ways. However, estimating two corrective filters, one for voiced and one for unvoiced sounds, would result in degradation of the corrected speech signals whenever a wrong voiced/unvoiced decision is made. Therefore, we only estimated one corrective filter.

First, the average power spectral density (psd) from the reference database r_p is estimated using a modified periodogram,

$$\mathcal{P}_{r_p}(f) = \frac{1}{||w||^2 K}\sum_{t=1}^{K} P_t^{(l)}(f) \tag{6}$$

where w is a hamming window, K is the total number of speech frames extracted from the reference database and $P_t^{(l)}(f)$ is given by

$$P_t^{(l)}(f) = |\sum_{n=0}^{N-1} w(n)\, s_t(n) \exp(-j2\pi f n)|^2 \tag{7}$$

where s_t is a speech frame from the lth observation sequence at time t.

Next, for the observations where hypothesis H_0 is rejected the average psd, $\mathcal{P}_{r_i}^{(l)}(f)$[2], is estimated in the same way as in Eq. (6). Observations from the same database and with similar likelihood scores are grouped together in a new observation set, and on this new set the average psd is estimated.

Lastly, the autocorrelation function, $\rho_{r_i}^{(l)}(\tau)$, is estimated as

$$\rho_{r_i}^{(l)}(\tau) = \int_{-1/2}^{1/2} \left(\mathcal{P}_{r_p}(f) - \mathcal{P}_{r_i}^{(l)}(f)\right) exp(j2\pi f \tau) df \tag{8}$$

[1] We assume that the sample size of each observation sequence is large so that the t distribution can be approximated by the standard z distribution

[2] Reminder: $\mathcal{P}_{r_i}^{(l)}(f)$ denotes the average power spectral density of the lth sequence from the recording session r_i

Given the autocorrelation samples $\rho_{r_i}^{(l)}[k]$ for $k = 0, 1, \ldots, q$, the coefficients of an AR (*autoregressive*) corrective filter of order q may be determined by solving a set of q linear equations (*Yule-Walker equations*) [6]. Filtering the speech signal from the $r_i^{(l)}$ session with the obtained AR corrective filter, the distance between the psd of reference and the psd of the filtered data, $\hat{\mathbf{O}}_{r_i}^{(.)}$, is decreased while the likelihood of the filtered data, $\mathcal{L}(\hat{\mathbf{O}}_{r_i}^{(.)}|\Lambda_{r_p})$, is increased. In all of our experiments, the filtered speech data was judged in informal listening tests as having the same voice quality as that of the reference.

4. OBJECTIVE EVALUATION

Because the statistic decision of rejecting or accepting the hypothesis H_0 has been based on the likelihood that the model Λ_{r_p} has produced the observations from the recordings r_i, we picked the obvious choice: as measure of the effectiveness of the compensation method we use the values of the log-likelihood function of the filtered data for the model Λ_{r_p}: $\mathcal{L}(\hat{\mathbf{O}}_{r_i}^{(m)}|\Lambda_{r_p})$, where $\hat{\mathbf{O}}_{r_i}^{(m)}$ is the mth observation sequence from the filtered part of the r_i session.

5. IMPLEMENTATION

The voice quality compensation system is based on the use of GMM and of AR corrective filters. In this section, practical issues on the estimation of GMM parameters, on the order of AR filters and on the implementation of the system, are described.

The parameters of the GMM are estimated by the EM algorithm. In the present work, the GMM parameters are initialized by a standard binary splitting *Vector Quantization*, (VQ), procedure: the weight, mean vector and covariance matrix of each component are estimated independently using the clusters obtained by VQ of the vectors extracted from the reference database. We decided to use diagonal covariance matrices for the GMM, in effect assuming statistical independence between each extracted vector. The speaker of the database was modeled by a 64 component GMM. The model was trained using 16-dimensional mel-cepstral vectors. It is important to note that in contrast with compensation methods in a speaker recognition task, where feature vectors are normalized by removing the long-term mean of the vectors prior to modeling by GMM, the vectors used here are not normalized. This is because we want first to detect any kind of "mismatch" between the reference and the testing database and then remove it; it is not desirable for our application to "blindly" normalize the feature vectors and then try to detect variabilities in voice quality. Another difference with the compensation methods used in speaker recognition is that the first cepstrum coefficient has been included into the training and testing process. Again, the main reason is that this coefficient carries useful information in our task i.e., the level of a recording session. The mel-cepstral data were extracted from non-overlapped segments. The sampling frequency used was 16kHz. The analysis frame (and rate) was $10msec$ while low-energy segments were excluded from the analysis. The size of the Fast Fourier Transform (FFT) for the estimation of power spectral densities was set to 256 and the order, q, of the AR corrective filters was 5.

6. RESULTS AND DISCUSSION

The voice quality detection and compensation system presented here was tested on a task of comparing 3 recording sessions: r_1, r_2, and r_3. The text used for these recordings was from the Wall Street Journal. Because of the same domain of the text, it was desired to have the same quality of voice for all recordings. The sampling frequency for the speech files was 16kHz. The length in minutes of each recording session, when small energy (e.g., silence) signals are excluded, was: r_1: 68 minutes, r_2: 118 minutes and r_3: 60 minutes. In order to select the reference database we decided to build a GMM in each of these sessions and measure their intra-session variability. The session with the smallest variance in log-likelihood is considered to be of *accepted* quality and was selected as reference. In addition to this criterion, an experienced listener (with experience in testing voice quality for speech synthesis) was asked to check samples from the previously selected recording session to assure that the quality of speech was acceptable. Each GMM was trained using approximately $60,000$ 16-dimensional mel-cepstral vectors (10 minutes) extracted from the beginning of each session (no fatigue effects were expected). To measure intra-session variability, each of the sessions was segmented into 3 minutes segments and the average log-likelihood in each segment was estimated. Fig. 1 shows the results for the three recording sessions. Note that as the first segments have been used for GMM training, the log-likelihood for these segments will be higher compared to the other segments. Also for plotting purposes, only 20 segments from each session were used, but covering the complete session (from the start until the end of the session). The

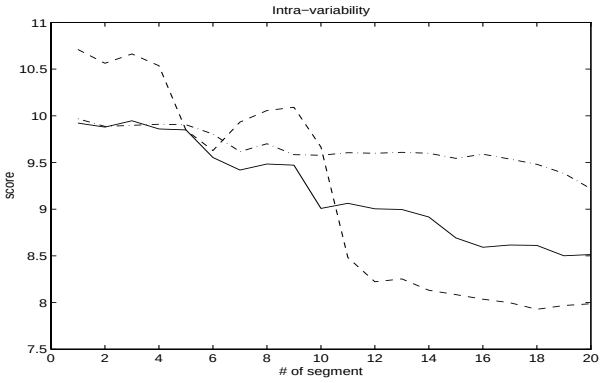

Figure 1: Intra-variability of recording sessions: r_1 solid line, r_2 dashed line and r_3 dash-dotted line.

scores shown in Fig. 1 are the average log-likelihood from each segment with reference model Λ_{r_i}. The functions $\mathcal{L}(\mathbf{O}_{r_1}^{(l)}|\Lambda_{r_1})$, $\mathcal{L}(\mathbf{O}_{r_2}^{(l)}|\Lambda_{r_2})$ and $\mathcal{L}(\mathbf{O}_{r_3}^{(l)}|\Lambda_{r_3})$ are represented by a solid line, a dashed-line and a dash-dotted line, respectively. For comparison, the score from a different speaker (of the same gender, female, and approximately the same range of fundamental frequency values) using any model of Λ_{r_i}, $(i = 1, 2, 3)$, was found to be around 4. Fig. 1 also shows that the recording session with the smaller variability was r_3. Therefore, it was selected as the reference session, r_p. As discussed earlier, the intra-session variability shown in Fig. 1 can be explained by effects of the duration of the recordings (e.g., fatigue after many hours of recordings) or by slightly different positions of the microphone during the recording. The

other two sessions have bigger variance than r_3 with r_2 having the largest. In fact, r_2 was the longest recording session.

Fig. 2 shows the log-likelihood functions $\mathcal{L}(\mathbf{O}_{r_i}^{(l)}|\Lambda_{r_p})$ using a solid line for $i = 1$, a dashed line for $i = 2$ and a dash-dotted line for $i = 3$ ($p = 3$). In the same figure the confidence interval of 99% for the reference session, r_3, is represented by the two straight solid lines. An important point to note is that r_2 has a significant lower average log-likelihood score compared to those of the other two recordings. After investigating this point, it was discovered that during r_2 the pre-amplifier used had malfunctioned, affecting the recording in a similar way as a pre-emphasis filter does. It has also reported that during the first recording, r_1, the position of the head-mounted microphone had to be corrected quite often. The overall average log-likelihood per session was: for r_1, 9.0974, for r_2, 8.0646 and for r_3, 9.6506. The voice quality difference between r_2 and r_3 (or r_1) has been noted by several listeners. On the other hand, it was not possible to hear any difference between r_3 and r_1. These findings are in agreement with the log-likelihood scores. The segments with a log-likelihood score

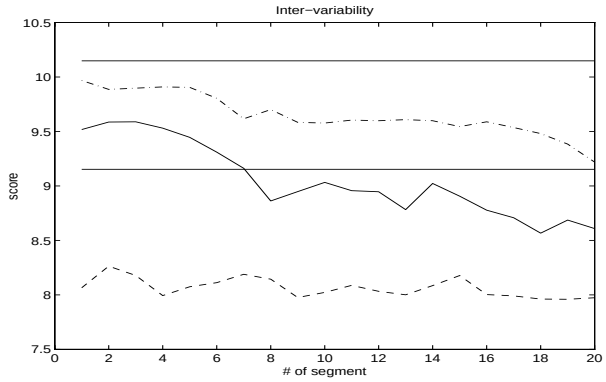

Figure 2: Inter-variability of recording sessions: r_1 solid line, r_2 dashed line and r_3 dash-dotted line.

smaller than the lower limit of the confidence interval were corrected using the technique described in Section.3. Fig. 3 shows the log-likelihood scores, $\{\mathcal{L}(\hat{\mathbf{O}}_{r_i}^{(l)}|\Lambda_{r_p})\}_{i=1,2}$, after correction. The average log-likelihoods after correction was 9.2654 and 9.2160 for r_1 and r_2, respectively. This is an improvement of 1.84% for r_1 and 14.27% for r_2.

An informal listening test was carried out using 5 listeners who were familiar with the speaker's voice (members of our TTS group at AT&T) and with experience in voice quality assessments. In addition, we used 5 other listeners with experience in assessments of speech coding quality, but unfamiliar with the speaker's voice. All listeners were able to detect the voice quality difference between the session r_2 and the two other sessions r_1 and r_3 before compensation, while after compensation no difference was detected. Again, listeners were unable to detect differences between r_1 and r_3 before or after compensation (for the corrected segments of r_1). It is also important to report that listeners did not notice any degradation of quality due to the correction process. The system described here is now an integrated component of our recording paradigm, insuring a consistent voice quality.

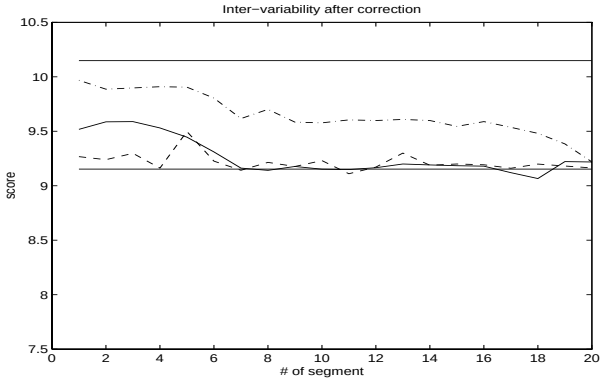

Figure 3: Inter-variability of recording sessions: r_1 solid line, r_2 dashed line and r_3 dash-dotted line.

7. CONCLUSION

In this paper we presented a system for automatic voice quality assessment and compensation. The method is based on Gaussian Mixture Models and AR filtering. Given a set of different recording sessions of the same speaker (widely now used for text-to-speech systems based on concatenation of acoustic units extracted from large databases), the proposed system is able to automatically select a recording session as reference, set statistical decision thresholds for compensation, and correct the segments from the recordings where needed. The quality of speech signals after correction is high. Results from subjective listening tests are correlated with the decisions that the system makes automatically. The effectiveness of the system was verified by estimating the post-correction likelihoods and by corresponding listening tests.

8. REFERENCES

[1] W. N. Campbell and A. Black, "Prosody and the selection of source units for concatenative synthesis," in *Progress in Speech Synthesis* (R. V. Santen, R.Sproat, J.Hirschberg, and J.Olive, eds.), pp. 279–292, Springer Verlag, 1996.

[2] K. Takeda, K. Abe, and Y. Sagisaka, "On the basic scheme and algorithms in non-uniform unit speech synthesis," in *Talking Machines* (G. Bailly and C. Benoit, eds.), pp. 93–105, North Holland, 1992.

[3] L. Neumeyer and M. Weintraub, "Probabilistic optimum filtering for robust speech recognition.," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, (Adelaide, Australia), pp. 417–420, 1994.

[4] D. A. Reynolds, *A gaussian mixture modeling approach to text-independent speaker identification*. PhD thesis, Georgia Institute of Technology, Aug. 1992.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc. Ser. B (methodological)*, vol. 39, no. 1, pp. 1–22 and 22–38 (discussion), 1977.

[6] S. M. Kay, *Fundamentals of statistical signal processing: Estimation theory*. PH signal processing series, Prentice-Hall, 1993.