Analysis on the Importance of Short-Term Speech Parameterizations for Emotional Statistical Parametric Speech Synthesis

Ranniery Maia  
Cambridge Research Laboratory  
Toshiba Research Europe Ltd, Cambridge, UK  
ranniery.maia@crl.toshiba.co.uk

Masami Akamine  
Corporate Research and Development Center  
Toshiba Corporation, Kawasaki, Japan  
masa.akamine@toshiba.co.jp

Abstract

This paper presents an analysis procedure to check on the connection between some commonly used spectral parameters for HMM-based speech synthesis, such as mel-cepstrum and mel-line spectral pairs, and excitation parameters related to both the pulse-noise balance as well as the glottal pulse shape itself, with the speaker’s emotion. The analysis is done with a labeled parallel emotional corpora. Firstly, the short-term parameters are analyzed through an emotion classification task, performed by two methods: $K$-means clustering and Gaussian mixture model (GMM)-based emotion classification. The best spectral and excitation parameterization in the sense of connection with the speaker’s emotion are then utilized to train emotion-dependent statistical parametric synthesizers. Finally, subjective tests of similarity are performed to assess the impact of the parameters.

Some studies have reported on the connection of glottal source parameters with the speaker identity [2], the speaker’s phonation style [3], and the speaker’s emotion [4, 5]. These analyses, however, have been mostly based on the Liljencrants-Fant (LF) model [6] or the normalized amplitude quotient [7]. As for the short-term spectrum, in [8] a cluster analysis indicates that the spectral envelope changes when speech goes from neutral to sad or happy states, and in [9] a spectral modification approach for unit concatenation-base expressive speech synthesis is proposed. These works indicate that there is indeed a connection between short-term spectral and excitation parameters and the speaker’s emotion.

1. Introduction

The research on expressive speech synthesis has gained considerable importance in the last few years. In most of the applications controllability of the expression of the synthetic speech according to the user’s intentions or feedback is desirable. To achieve this, some text tags can be used to represent this information. These tags are eventually conveyed to the frontend of a text-to-speech (TTS) system to synthesize speech with the desired style.

The synthesis of the speech signal with a desired style or expression can be performed through the manipulation of prosodic, spectral and excitation parameters. It is a common understanding that long-term prosodic information have a significant effect on the emotion of the synthesized speech. However, when it comes to hidden Markov model (HMM)-based speech synthesizers [1] the question to be asked then is how short-term spectral and/or excitation parameters, the commonly used features which compose the observation vectors of the HMMs, affect the style or expression of the synthesized speech.

2. Analysis of emotional short-term speech parameters

The analysis is performed by an emotion classification task, performed by both $K$-means clustering and GMM emotion identification. These tasks use as features the analyzed short-term spectral and excitation parameters ex-
tracted from an labeled emotional database.

2.1. The classification methods

2.1.1. K-means emotion clustering

Assuming that \( X = \{ x_0, \ldots, x_T \} \) corresponds to short-term speech parameters extracted from the entire emotional data (including all the emotions), \( X \) is split into \( K \) clusters \( \Omega = \{ \Omega_0, \ldots, \Omega_{K-1} \} \) so as to minimize

\[
D = \arg \min_{\Omega} \sum_{k=0}^{K-1} \sum_{x_t \in \Omega_k} \| x_t - \mu_k \|^2,
\]

where \( \mu_k \) is the centroid in \( \Omega_k \). Since the intention is to verify how the speech features represented by \( X \) are dependent on the speaker's emotion, the number of clusters \( K \) is set to the number of emotional classes \( C = \{ C_0, \ldots, C_{K-1} \} \). The performance of the clustering is measured by the normalized mutual information [10],

\[
U(\Omega, C) = \frac{I(\Omega, C)}{H(\Omega) + H(C)},
\]

where \( I(\Omega, C), H(\Omega) \) and \( H(C) \) are respectively the mutual information and entropies of \( \Omega \) and \( C \).

2.1.2. GMM-based emotion identification

In GMM modeling [11] the likelihood that a given feature vector \( x_t \) belongs to a class \( C_k \) is

\[
p(x_t \mid C_k) = \sum_{j=0}^{J-1} w_j N(\mu_j, \Sigma_j),
\]

where \( \mu_j \) and \( \Sigma_j \) are respectively the mean vector and covariance matrix of the \( j \)-th Gaussian component, \( w_j \) is its respective weight, and \( J \) is the number of components. In this work each emotional data was separately used to train an \textit{emotion dependent} GMM using the expectation maximization algorithm. After that, a given emotional test set \( Y = \{ y_0, \ldots, y_{T-1} \} \) had its class identified by

\[
\hat{C} = \arg \max_k \sum_{t=0}^{T-1} \log p(y_t \mid C_k).
\]

2.2. Speech parameters utilized in the analysis

2.2.1. Spectral parameters

Two types of spectral parameters were investigated: the mel-cepstrum and the mel- linear spectral pairs (mel-LSP). Mel-cepstrum and mel- linear prediction coefficients (mel-LPC) were computed from short-term periodograms of the speech signal through the mel-generalized cepstral analysis described in [12], with the conditions \( (\alpha, \gamma) = (0.42, 0) \) and \( (\alpha, \gamma) = (0.42, -1) \), respectively, where \( \alpha \) and \( \gamma \) are the warping frequency and pole-zero controlling factors, respectively. As a last step, mel-LPC were transformed into mel-LSP.

2.2.2. Excitation parameters

Mixed excitation approaches have been commonly used in statistical parametric synthesizers [1]. To create a mixed excitation signal, normally two procedures are taken into account: (1) the modeling of the glottal pulse shape; (2) the mixing of pulse and noise. Usually, different parameterizations are used separately for each one of these procedures.

Two parameters related to glottal pulse modeling were analyzed. The first one corresponded to the anti-causal part of the complex cepstrum calculated from interpolated amplitude and unwrapped phase spectra [13]

\[
\phi(n) = -\hat{x}(-n - 1), \quad n = 0, \ldots, C - 1,
\]

where \( C \) is the order of the phase parameters, and \( \hat{x}(n) \) is the complex cepstrum. The other glottal pulse feature is based on the group delay, similar to the approach described in [14]

\[
\nu(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \tau(\omega) e^{i\nu} d\omega, \quad n = 0, \ldots, C - 1, \quad (6)
\]

with \( \tau(\omega) \) being the group delay and \( C_\nu \), the size of \( \nu(n) \).

For parameters that control the mix of pulse and noise, the band-aperiodicity coefficients given by

\[
b(n) = \sum_{\omega l \in B_n} a(\omega l) / B_n, \quad (7)
\]

were used, where \( a(\omega l) \) is the aperiodicity measure at frequency \( \omega l \) [15], \( B_n \) is the \( n \)-th bandwidth, and \( b(n) \) is the corresponding band-aperiodicity coefficient.

3. Experimental results

3.1. Speech corpora and parameter extraction

A North-American English emotional database uttered by a female speaker was utilized in the current analysis. The emotion classes considered were: \( C_0 = \text{angry} \), \( C_1 = \text{happy} \), \( C_2 = \text{neutral} \), \( C_3 = \text{sad} \). The data comprised 609 sentences in each emotion.

The data was segmented at the phone level using hidden Markov models (HMMs), where each HMM state emission corresponded to a 5-ms duration feature vector.

To extract spectral and excitation parameters pitch-synchronous Fourier analysis was conducted on the speech signals. After that, frame-based amplitude and unwrapped phase spectra were obtained as described in [13]. From the amplitude spectra 40 mel-cepstral coefficients, including the 0-th term, and 39 mel-LSPs plus the log power term were computed. From the phase spectra 39 group delay features were obtained. 39-order complex cepstra were also calculated from both amplitude and phase spectra. Finally, 22 bark [16] band-aperiodicity features were extracted from the aperiodicity parameters [15].
3.2. $K$-means clustering results

Figure 1 shows the normalized mutual information for each phone of the $K = 4$ clusters formed after the clustering process. The initial centroids were randomly chosen from the data points. Table 1 summarize the results as the average across the phones. It can be noticed the spectral features mel-cepstrum and mel-LSP are more appropriate for this classification task than the excitation ones. This implies that mel-cepstrum and mel-LSP are thus the most emotion-dependent features.

3.3. GMM classification results

Figure 2 shows the identification accuracy rate for each phone according to each feature. Table 2 shows the overall results as the average of the accuracy across all the phones. Among the spectral features, the mel-LSP parameters achieved the highest emotion identification rate, which implies that they are the most emotion-dependent among the tested features. Among the excitation features it is not clear which ones are the best.

3.4. Experiments with synthetic speech

Four emotion-dependent HMM synthesizers were trained using the emotional corpora described in Section 3.1. To test the impact of the emotional spectral and excitation parameters on the expressiveness, listening tests were conducted through a web-recruitment system, with a minimum requirement of three judgments per test utterance. A total of 31 test sentences were used. In average, 85 subjects took part in each test. The subjects were asked to compare synthesized versions of a given sentence with a natural emotional version and rate the similarity from 1 to 5. Based on the results of Section 3.3, mel-LSP parameters were chosen as spectral features, and complex cepstrum-based phase features were selected as the glottal pulse parameters. The latter ones were chosen because they have been shown to be effective in terms of improving synthesized speech quality [13]. Finally, band-aperiodicity parameters were also used to control the mix between glottal pulse and noise in the excitation signal. Each HMM observation vector was composed of the following streams: stream 1: log power, delta and delta-delta; stream 2: 39 mel-LSPs, delta and delta-delta; streams 3,4,5: $\ln F_0$, $\Delta \ln F_0$, $\Delta \Delta \ln F_0$, respectively; stream 6: 22 band-aperiodicity features, delta and delta-delta; stream 7: 39 complex cepstrum-based phase features, delta and delta-delta.

The test sentences were synthesized from parameters generated from a mix of streams from the emotional systems. In order to conduct this stream blending, the parameters modeled by the HMMs were tagged as follows: prosody: log power, $\ln F_0$ and duration; spectrum: mel-LSP: excitation: complex cepstrum-based phase features and band-aperiodicity.
shown that short-term spectral envelope is indeed important for enhancing expressiveness of the synthetic speech. Although the literature has shown that excitation features can also be associated with the speaker’s emotion, three possible reasons for the failure of replicating such results here are: (1) the use of a small database; (2) the use of glottal features encapsulated in the cepstral domain; (3) the failure of accurately representing short-term glottal information (pitch marking errors, phase unwrapping inaccuracy). In the future we intend to solve these issues and provide a more accurate analysis of glottal pulse features and emotion for HMM-based synthesizers.

5. References


