Automatically Creating a Diphone Set from a Speech Database

Thomas Ewender and Beat Pfister

Speech Processing Group
Computer Engineering and Networks Laboratory
ETH Zurich, Switzerland
{ewender,pfister}@tik.ee.ethz.ch

Abstract

This paper presents a measure that scores various aspects of phone quality. The measure is designed to penalize phone instances with one or several characteristics that are not desirable in concatenation-based speech synthesis. Depending on the phone type, these aspects amongst others include spectrum, phase, fundamental frequency, duration, voicing and plosive quality. We applied this quality measure to select diphone sets from four different speech databases and demonstrate the quality of these diphone sets by means of synthesis examples. The quality of these examples showed that the proposed measure can be applied to select a high-quality diphone set from a speech database.

Index Terms: speech synthesis, phone quality measure

1. Introduction

In concatenative speech synthesis, corpus generation still involves tedious manual or semi-automatic selection of units. In diphone synthesis, the segments for a diphone set are either selected manually from a speech database or extracted from designated diphone carrier words with one or two diphones embedded in one carrier word (see [1]). In the same way, creating a unit selection voice involves manual work in the postprocessing of speech recordings to identify low quality segments, labeling errors or pronunciation variants.

This demand for manual work is due to the lack of a quality measure that could help to decide which phone segments are appropriate to be selected. Automatic phone quality judgement for corpus creation was only considered to a small extent so far. In [2] the best diphone variant is selected using the cepstral distance between the two semi-diphones and the corresponding phone centroids as the only automatic measure. Unit selection does not directly consider the quality of the selected units in their target costs during synthesis, because no acoustic properties for the target units are known. Phone quality is considered only indirectly through the concatenation costs (see [3]), which are high for spectral discontinuities. Various measures to detect these discontinuities were proposed in [4] and [5].

In this paper we present a phone quality measure that has been designed to automatically select diphones from a speech database. For each diphone the following criteria are important:

1. The two involved phones must be heard as clearly articulated and unambiguously identifiable instances of these phones.

2. The signal of the phones has to be suitable for prosodic modification (e.g. by means of PSOLA, see [6]) without impairing the perceived speech quality.

3. The chosen diphones can be concatenated without audible artifacts at the concatenation points.

Our phone quality measure can not only be used for diphone selection but for concatenation synthesis in general. We applied it here in the context of diphone synthesis because the high number of concatenation points immediately points to possible weaknesses of the method.

The phone characteristics assessed by the quality measure will be outlined in the next section. The phone quality measure as a combination of these characteristics will be described in Section 3. Section 4 will describe the application of this measure to the automatic extraction of a diphone set. Section 5 will describe our experiments and the results. Concluding remarks will be given in Section 6.

2. Phone quality aspects

Several aspects contribute to the overall subjective impression of the quality of phones. The set of aspects depends on the type of a phone, e.g. if the phone is voiced or unvoiced, or if it is plosive or stationary. In the following, the aspects that are relevant for the quality of phones, especially in the context of diphone synthesis, will be described. For each of these aspects we will present features to score that aspect. Since phone quality is assessed in the context of diphone selection, we want to assign no or a very small penalty value (less than 1) to a phone if the aspect under consideration is within a certain limit that is acceptable from a perceptual point of view, and a high penalty otherwise. This concept is reflected in the various power values that are used in the computation of these penalties. To obtain one value for the overall quality of a phone, the sum of these penalties is taken as described in Section 3.

Several aspects of phones depend on features such as duration, fundamental frequency (F_0) and pitch marks. The extraction of these features is not presented in this paper. For the segmentation of speech signals into phones we used the fully automatic method described in [7]. F_0 detection and pitch marking are described in [8] and [9]. For the F_0 detection a frame shift of 5 ms and a frame size of 50 ms was used.

2.1. Spectral characteristics

Spectral properties play a crucial role in assessing the quality of phones, in particular of stationary ones. In the context of diphone synthesis we are not only interested in excluding phone instances that are not unambiguously identifiable but also instances that are not typical for this phone as pronounced by a certain speaker. In other words, we need a score that penalizes inaccurately pronounced phones as well as untypical ones.

We score the spectral appropriateness of phone instances

using the cepstral distance between the frames of the considered phone instance and the corresponding phone centroid, i.e., an average typical cepstrum of that phone. This phone centroid is computed iteratively. First, the initial centroid is computed by averaging the cepstral vectors of all frames of all instances of the considered phone. Then the iteration is as follows: For each phone instance, a weighted sequence of five frames (covering a total length of 60 ms) with minimal cepstral distance to the current centroid is determined, and a new centroid is computed by taking the average of all these minimum-distance frame sequences. This step is repeated until convergence occurs or a maximum number of iterations is reached.

To describe the cepstra we have found MFCCs most suitable. As distance measure we used the Euclidean distance between the 12-dimensional cepstral vectors whereby the zeroth cepstral coefficient was neglected.

From listening experiments we found that phone instances with a distance $h < 8$ from their corresponding centroid still are perceived as very clear. Therefore we designed a function that strongly penalizes phones with higher distance values:

$$P(h) = \exp(0.4(h - 8)) \tag{1}$$

2.2. Phase characteristics

For some speakers it can be observed that instances of the same phone have very diverse waveforms although they sound similar and the cepstral distance between them is quite small. This diversity can be attributed to a considerable difference between the phases of these phones. If such phones happen to be concatenated, an artifact may be clearly audible from the resulting signal. An example of such a case is shown in Figure 1.

A similar effect may be caused by erroneous pitch marks. E.g. for a nearly sinusoidal speech signal it is not always clear if the pitch marks have to be set at the energy maxima that coincide with the positive or with the negative maxima of the fundamental wave (see also [8]).

Both these problems can be detected from the position of the pitch marks relative to the phase of the fundamental wave. If this phase value φ for a phone instance differs considerably from the average phase value μ_φ over all instances of that phone, one of the above mentioned cases applies and this phone instance should be penalized. With the phase values expressed in radians, we apply the following penalty function:

$$P(\varphi) = (3 \cdot (\varphi - \mu_\varphi))^4 \tag{2}$$

2.3. Fundamental frequency characteristics

If PSOLA-based fundamental frequency (F_0) and duration modification is applied to speech segments that are to be concatenated, the F_0 characteristics of these segments may cause several issues. If the F_0 at the end of one speech segment deviates considerably from the F_0 at the beginning of the next segment that is going to be concatenated, the degree or even the direction of F_0 modification required to realize a smooth contour changes abruptly at the concatenation point. Furthermore, speech segments with rapidly rising or falling F_0 are not suited to be transformed into segments with constant F_0 or even with an opposite direction of F_0 movement.

To account for these effects, we have defined two penalties. The first one penalizes phones with a fundamental frequency that considerably deviates from μ_f, i.e. from the mean value over all instances of that phone:

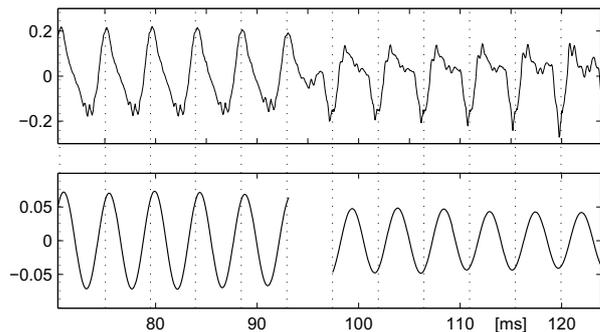$$P_1(F_0) = (10 \cdot |f - \mu_f|)^3 \tag{3}$$

Figure 1: *Artefact at around 95 ms in a speech signal resulting from the concatenation of the diphones [tʏ] and [ʏɐ] (top) and the corresponding fundamental waves (bottom). It can be seen that the pitch marks in [ʏ] of the left diphone have been set near the positive maximum of the fundamental wave, whereas they are near the minimum in [ʏ] of the second diphone.*

Note that the logarithm of F_0 is used, which makes the formula equally valid for male and female voices. Therefore, the mean F_0 value over all N frames of a phone instance is $f = \frac{1}{N}\sum_n (\log F_0(n))$.

The variation of F_0 within a phone instance is expressed with the temporal derivative and results in the penalty

$$P_2(F_0) = \exp(200(f' - \mu_{f'} + \sigma_{f'})) + \exp(100(\hat{f}' - \mu_{\hat{f}'} + \sigma_{\hat{f}'})), \tag{4}$$

where $f' = \frac{1}{N}\sum_n |\log F_0(n) - \log F_0(n-1)|$ is the mean absolute derivative of the F_0 of this phone instance and $\mu_{f'}$ and $\sigma_{f'}$ are the mean and standard deviation of f' over all instances of that phone. Similarly, $\hat{f}'$ is the mean over the highest 25 % of the components that contribute to f', and $\mu_{\hat{f}'}$ and $\sigma_{\hat{f}'}$ are the mean and standard deviation of $\hat{f}'$ over all instances of that phone.

2.4. Duration characteristic

In concatenation synthesis, longer phones are preferred over shorter ones since, generally, shortening impairs the quality much less than lengthening. However, in automatically segmented speech signals, phone instance durations that are much higher than the average phone duration may originate from segmentation errors. The following function accounts for both of these aspects:

$$P(d) = 10 \cdot |\log d - (\mu_d + \sigma_d)|^3 \tag{5}$$

This function prefers phones that are one standard deviation σ_d longer than the mean duration value μ_d. Note that durations are in seconds and are used in the log domain, i.e. the mean duration of J phone instances is $\mu_d = \frac{1}{J}\sum_j \log d_j$ with $j = 1 \ldots J$.

2.5. Voicing characteristics

In the context of speech synthesis, voicing has two aspects. First, breathy vowels are not desirable because they do not sound clear, and second, irregularly pitched speech is problematic for prosodic modification with PSOLA. Therefore, we want to penalize voiced stationary phones with these properties. As features we used the output of the frame classifier presented in [9], that decides if speech frames are voiced, unvoiced

or mixed and distinguishes between regularly and irregularly pitched frames. The number of mixed frames N_m and the number of irregularly pitched frames N_i of a phone with N frames are considered in the penalty function as follows:

$$P(v) = 20 \frac{N_m + N_i}{N} \qquad (6)$$

2.6. Characteristics of plosives

The most important characteristic of plosives is the burst, i.e. the sudden air flow after the release of the closure. The burst has to be strong enough to be clearly identified. This can easily be detected from the short-term power of the speech signal.

In some languages (e.g. in German), correct pronunciation requires to articulate unvoiced plosives either with or without aspiration, depending on the context. Therefore we created a method to decide from the speech signal which plosives are aspirated and which ones are not. We used this information to correct the labels of aspirated and non aspirated plosives if necessary.

From phonetics it is known (see e.g. [10]) that unvoiced plosives with a voice onset time (VOT) greater than some 50 ms are clearly heard as aspirated, whereas no aspiration is heard if the VOT is less than 20 ms. To assess the aspiration of unvoiced plosives we have to detect the release point and the start of voicing. This is illustrated in Figure 2. The release point, which is the boundary between the closure and the burst, is determined by looking for the point of maximum increase of the energy in the band from 2 to 8 kHz (see [11, 7]).

The start of voicing is detected from the intensity of the fundamental wave. The fundamental wave for a sample i of a signal $x(\cdot)$ is computed as follows:

$$f(i) = \frac{\sum_j x(j) \cdot w(j - i)}{\sum_j w(j)}, \qquad (7)$$

where $w(\cdot)$ is a Hamming window of length $T_0 = 1/F_0$. The convolution of the speech signal with a Hamming window of length T_0 corresponds to a low-pass filter with a cut-off frequency of F_0 and zeros at the harmonics. Because a T_0 contour as it results from the optimization described in [9] is specified with one value per frame, it has to be interpolated to obtain a T_0 value for each sample of the speech signal. The intensity of fundamental wave is then computed as follows:

$$e(i) = \sqrt{\frac{\sum_j [f(j) \cdot u(j - i)]^2}{\sum_j u(j)}}, \qquad (8)$$

where $u(\cdot)$ is a Hamming window of length $2T_0$ centered at 0. From this intensity curve the start of voicing is detected by means of a threshold that depends on an estimate of the speech loudness. Finally, the difference between the start of voicing and the release point yields the required VOT.

2.7. Signal intensity

In natural speech, the signal intensity varies considerably even between instances of the same phone. In order to avoid that concatenated diphones produce phones with very different intensity in the first and the second half, we penalize stationary phones with a short-term signal intensity g that differs more than 6 dB from the average intensity μ_g over all instances of this phone:

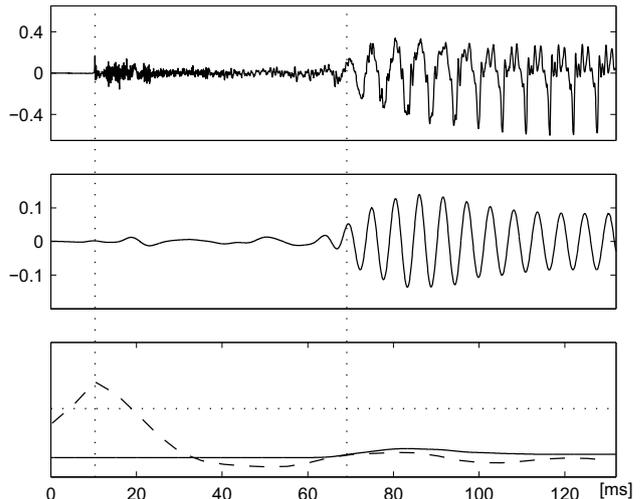$$P(g) = (0.2 \cdot (g - \mu_g))^4 \qquad (9)$$

Figure 2: *Estimation of the VOT in a speech signal (top plot) with the phones [taɪ]. The fundamental wave is shown in the middle plot. The maximum energy change curve (dashed, bottom plot) defines the release point at 10 ms. The intensity of the fundamental wave (bottom plot) crosses the threshold at 68 ms. Therefore, the VOT is greater than 50 ms and the unvoiced plosive is considered to be aspirated.*

3. Phone quality measure

From the above described scores that penalize various phone aspects individually, an overall score has to be derived. As already mentioned, the set of aspects to be applied depends on the type of phone. This is not a problem, because in the context of diphone selection we only have to compare instances of the same phone and not arbitrary phones. Furthermore, the overall score does not have to represent an absolute or even interpretable value. The overall score is only needed to rank instances of the same phone.

As overall score we used the sum of the penalties resulting from the functions given in Section 2. No normalization of these penalties was applied as the penalty functions are designed in such a way that the limits of acceptability for each aspect that the penalty functions describe range around the same value.

For the development of the penalty functions we used an interactive tool, which allows diphone instances to be selected from a ranked list and to be used in synthesis. Each diphone can be played in different contexts to subjectively judge its quality, not absolutely but only with regard to the rank order. In this way, we were able to identify aspects that strongly influence synthesis quality and therefore had to be integrated into the phone quality measure.

4. Automatic extraction of a diphone set

Given a speech database, i.e. a sufficiently large collection of recorded sentences and the corresponding text, the automatic process for diphone extraction comprises the following steps:

1. Phonetic transcription: The phonetic transcription of the text is generated with the SVOX speech synthesizer. This synthesizer allows for various types of outputs, amongst others a phonological transcript that includes the phonetic transcription of the words augmented with abstract prosodic information such as syllable stress level and phrase boundaries.

2. Segmentation into phones: Based on the phonetic transcription, a fully automatic HMM-based segmentation of the speech signals is performed as described in [7].

3. Definition of diphone set: The list of all phone transitions in the recordings is extracted from the segmentation. Note that this list cannot be compiled from the phonetic transcription, because only after the segmentation we know which words are separated by pauses. In these cases there is no direct transition from the last phone of a word to the first phone of the next word.

4. Computation of phone scores: The scores of all phone segments are computed as described above.

5. Computation of diphone scores: The score of a diphone is based on the score of the respective phone instances. We used the sum of the phone scores as a diphone score.

6. Setting diphone boundaries: A diphone boundary, which should be somewhere in the middle of a phone, is defined as the point with minimal cepstral distance from the corresponding phone centroid. We have found that for robustness reasons the distance measure has to consider several weighted frames. This method is applicable only for stationary phones. For plosives the diphone boundary is set right before the release point.

Finally the best-scored instance of each diphone is extracted from the speech signals.

5. Evaluation

We applied our diphone extraction method to four speech databases: two English and two German, where for both languages there was a male and a female one. These databases contained sentences of various length. The overall length of the speech signals were 150 minutes for the female German and the male English database, 85 minutes for the female English and 45 minutes for the male German one. From these four databases we created diphone sets in a fully automatic way as described in Section 4.

In order to assess the quality of the resulting diphone sets, we used them to synthesize example sentences. In order to exclude possible artifacts in the synthesized speech signal that may originate from weaknesses of other components of the synthesis system, e.g. from prosody control, we synthesized the example sentences as follows. We selected a small set of sentences from each of the four databases. Note that these sentences were excluded from the above described diphone extraction. From these sentences we extracted the prosody, i.e. the durations and the F_0 values of the phones. This allowed us to use diphone concatenation to generate synthetic speech with natural prosody.

The results from this experiment were as follows: The example sentences produced with the female German and with the female English diphone sets showed very little distortion and sounded quite natural. More distortions were audible in some of the examples from the male English diphone set, others were virtually free of defects. The example sentences from the male German diphone set showed more defects than those from the other three diphone sets, possibly because the size of the male German database, which contains only 45 minutes of speech, is rather limited.

We demonstrate our results by means of examples on the web at http://www.tik.ee.ethz.ch/spr/test_sentences/.

6. Conclusion

The quality of the synthesis examples shows that our proposed phone quality measure can be applied to select a high-quality diphone set from a speech database. As a consequence, tedious manual work in the creation of such diphone sets can largely be eliminated. We believe that other synthesis methods can also benefit from our phone quality measure. In unit selection, the phone quality measure may be used as a criterion in database pruning to reduce the size of the system, and in the selection step as an additional feature for the target costs of candidate units.

One limitation of our approach is that the penalty functions for the phone characteristics are motivated by acoustic inspection, and in addition taking the sum of the penalties may not entirely represent the perceptual impression. Future research could aim at replacing these penalty functions by a machine learning approach to weight the features and combine them in a non-linear way. However, the approach to create an appropriate data set for the training of this classifier is not obvious.

8. References

[1] K. Lenzo and A. Black, "Diphone collection and synthesis," in *Proceedings of ICSLP*, 2000.

[2] C. Traber, K. Huber, et al., "From multilingual to polyglot speech synthesis," in *Proceedings of Eurospeech'99*, Budapest, September 1999, pp. 835–838.

[3] N. Campbell and A. Black, *Prosody and the selection of source units for concatenative synthesis.* Springer, 1996, ch. 22, pp. 279 – 292.

[4] Y. Stylianou and A. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proceedings of ICASSP*, 2001.

[5] R. Donovan, "A new distance measure for costing spectral discontinuities in concatenative speech synthesisers," in *ISCA Tutorial and Research Workshop (ITRW)*, 2001.

[6] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453 – 467, December 1990.

[7] S. Hoffmann and B. Pfister, "Fully automatic segmentation for prosodic speech corpora," in *Proceedings of Interspeech*, Makuhari (Japan), September 2010, pp. 1389–1392.

[8] T. Ewender and B. Pfister, "Accurate pitch marking for prosodic modification of speech segments," in *Proceedings of Interspeech*, Makuhari (Japan), September 2010, pp. 178–181.

[9] T. Ewender, S. Hoffmann, and B. Pfister, "Nearly perfect detection of continuous F0 contour and frame classification for TTS synthesis," in *Proceedings of Interspeech*, Brighton, September 2009, pp. 100–103.

[10] L. Lisker, "Cross-Language Study of Voicing in Initial Stops," *JASA*, vol. 35, pp. 384–422, 1963.

[11] L. Golipour and D. O'Shaughnessy, "A new approach for phoneme segmentation of speech signals," in *Proceedings of Interspeech*, 2007, pp. 1933–1936.