Two Features to Check Phonetic Transcriptions in Text to Speech Systems

Stefano Sandri, Enrico Zovato

Voice Technology
Loquendo, Vocal Technology and Services
Turin, Italy
{stefano.sandri,enrico.zovato}@loquendo.com

Abstract

The paper describes a framework to overcome some problems in the analysis of speech corpora used in text-to-speech systems. In particular two kinds of errors that can produce disagreeable effect at synthesis level have been examined. The first of them is the incorrect transcription of pauses (and more generally low energy intervals) and the second one is the mismatch between voiced intervals and the phonetic symbol that should represent them. For the first problem a statistical approach has been used, by comparing some features of the detected low energy intervals (LE) with those of trained data. The second problem has been faced extracting the voiced/unvoiced intervals (VU) and checking the coherence with the phonetic transcription and segmentation.

1. Introduction

Corpus based text-to-speech systems are reaching high quality levels thanks to the storage and computation capabilities available nowadays. Very large speech databases can be involved in these systems so that a good statistical coverage of different phonetic and prosodic contexts is reached. On the other hand such material has to be correctly analysed, and in particular it has to be phonetically segmented and labeled. Manual operations would require a very large effort in terms of time and people involved. This is the reason why our text-to-speech system Actor® [1] (once Cselt, now Loquendo multi-voice and multi-language synthesis system) relies on an automatic phonetic aligner, based on Markov models [2],[3].

Although such automatic analysis is rather reliable, still some mismatches between signal and transcription may occur. Among the possible causes the three most probable are: (1) the speaker has not pronounced the text correctly, (2) the phonetic transcriber has made a mistake, (3) the aligner has not worked properly due to an insufficient statistical coverage of the material used to train the acoustic models. In any case such mismatches in the database can cause rough errors at synthesis level. This is the reason why manual corrections are always necessary to maintain high quality levels in speech synthesis. In order to reduce the manual effort required to control the database consistency, two features to automatically check the analysis data are proposed. They are low energy intervals (speech pauses, closures, glottal pauses, etc.) and voiced/unvoiced intervals. The output of this phase can be useful to correct the phonetic transcriptions, but can also be used in some cases, to suggest phonetic rules for the automatic phonetic transcriber.

In the next paragraph an overview of the analysis process that is needed to produce the database is provided. In the following two paragraphs the features check will be described and results and discussions will be reported in section 5.

2. Text-to-Speech database

Actor® text-to-speech system is based on a concatenative technology, that is, synthesis is produced linking variable length speech units depending on phonetic, syntactic and prosodic characteristics of the input text. The quality of this system is strongly dependent on the corpus size and on the statistical coverage of the collected speech material as well as on the accuracy of its phonetic and acoustic analysis.

The starting point of the analysis process consists of a set of high quality speech signals acquired in a recording studio and produced by professional speakers. The texts that have to be recorded are designed in order to have a good phonetic and prosodic coverage. This is achieved by means of a statistical analysis of large text corpora and by the use of greedy algorithms to extract a minimal subset of sentences that provide the necessary coverage.

The phonetic transcription is originally obtained from the literal transcription by means of an automatic phonetic transcriber that is the same that will be used at synthesis level when text has to be converted into speech.

Figure 1. Analysis scheme including the check stage
This transcriber is based on large size lexicons and on phonetic rules. The phonetic aligner provides the temporal segmentation [3] (figure 1). Subsequently, diphone segmentations, pitch markers and prosodic features are calculated. In this way the database containing all the speech signals and the analysis data (there is a sort of index that refers to every phoneme of the acoustic material), is available for speech synthesis.

3. Low Energy intervals

The goal of this kind of analysis is to evaluate if the low energy intervals in the database speech signals are compatible with the given phonetic transcription and segmentation. These intervals include plosive closures, glottal pauses, some fricatives and speech pauses (that are labeled with a special symbol in the transcription). Sometimes the speaker doesn’t respect the pauses expected in the texts that have to be recorded; in other cases extra pauses could be inserted. These errors are not frequent, but if not detected can create severe problems both at analysis level, by confusing the automatic aligner, and at synthesis level, when a speech segment including an extra pause is selected. In fact the pauses present in the sentences are labeled with specific prosodic labels that are fundamental in the speech unit selection criteria.

To retrieve this kind of mismatch all the low energy intervals in the speech signals have to be detected and then compared with the phonetic transcription and segmentation. To make this control, an a priori knowledge of where LE intervals are expected is necessary; in practice it has to be known which phonetic contexts most probably contain a LE interval. A statistical approach has been used and for each phonetic context composed by two phones the probability of the second of them containing a low energy interval has been calculated.

To estimate these probabilities, a training procedure, based on a frequency model, has been provided. The probability associated to a given context (phoneme pair) has been defined as the ratio between the number of occurrences containing the LE interval and the total number of occurrences of that context:

\[ P_{xy} = P[p_{y}^{LE} | p_{x}, p_{y}] = \frac{N(p_{y}, p_{x}^{LE})}{N(p_{x}, p_{y})} \] (1)

Where: 
- \( P_{xy} \) is the probability that the context contains a low energy interval
- \( N(p_{y}, p_{x}^{LE}) \) is the number of occurrences of the context containing the LE interval
- \( N(p_{x}, p_{y}) \) is the number of occurrences of the context without the LE interval

3.1- Training algorithm

To calculate the context probabilities \( P_{xy} \), a significant set of speech signals and their transcriptions and segmentations are necessary. This set of data should:

• Provide good coverage of phonetic contexts containing LE intervals
• Have correct phonetic transcriptions
• Have correct alignments

Manually checked material is ideal in this phase, however if the phonetic transcriber works correctly and the automatic aligner performs high recognition rates (>90%), and if it is possible to make the hypotheses that the pause errors are not frequent (i.e. they provide little statistical noise), then the first output of the phonetic aligner could be used.

Speech signals have been analyzed by 20 ms long frames, shifted by 10 ms. For each frame the log energy and the zero crossing rate have been calculated. All the frames having the log energy below a particular energy threshold have been marked as low energy frames. The threshold is given by:

\[ thr = E_{\text{min}} + (E_{\text{max}} - E_{\text{min}}) \cdot s \] (2)

\( E_{\text{min}} \) and \( E_{\text{max}} \) are respectively the minimum and maximum log energy in the sentence, and \( s \) is an empirical sensitivity factor that in our experiments has been set to 0.5.

In this way for each speech signal a certain number of LE intervals has been found. Each of these intervals has been described in terms of:

• Phonetic context associated (phoneme pair)
• Duration
• Zero Crossing Rate

The first information is extracted from the phonetic transcription and segmentation. At the end of this training phase, for each phonetic context the pause probability, its mean length and its mean zero crossing rate are available.

3.2- LE Analysis

Starting from the reference data obtained by training, it is possible to perform the analysis of the material that has to be checked in order to find insertion errors, or in other words LE intervals in the signal not expected in the transcription. First a low energy intervals detector (the same used in the training phase) is applied to the signals, and for each LE interval, its zero crossing rate is calculated. Each interval is associated to the best fitting phoneme in terms of time boundaries. In practice the transcription phoneme that contains the LE interval or ensures the best time overlapping with it, is selected. When a phoneme is associated to the LE interval, the two phones context is also characterized. Now a comparison with the trained data can be achieved and to this end an error coefficient has been defined. This coefficient has been used to select the most probable errors that occur so that manual check can be focused on a small portion of the data. This parameter takes into account the context probability, the difference from the mean duration of the pause in that specific context and the difference from the mean zero crossing rate.

The error coefficient has been defined as:

\[ C_{\text{err}} = w_{1} \cdot (1 - P_{e}) + w_{2} \cdot \text{len}_\text{diff} + w_{3} \cdot \text{zcr}_\text{diff} \] (3)

Where \( P_{e} \) is the probability that the context contains a low energy interval, len_diff is the percentage length difference,
4. Voiced/Unvoiced intervals

The second feature that can help in finding macroscopic analysis errors is the voicing information. The problem is again a comparison between signal and its labelling data (transcription, segmentation), given an a priori knowledge. In fact before making this kind of analysis we have to know which phonemes are voiced or, in other terms, are periodic. Also in this case voicing depends on the phonetic context. Typical examples could be the /s/ /z/ pair in the Italian and Brazilian Portuguese, or the /t/, /d/ pair in the American and British English.

The automatic transcription system is based on lexical and general phonetic rules, and exceptions to them are always possible. When there is a wrong phonetic symbol the speech unit associated to it can be incorrectly selected during synthesis. Moreover the pitch markers computation will fail since, in our system, it is performed for speech intervals that are labeled as voiced phonemes. To avoid this problem a tool to check the coherence between voiced intervals and their transcriptions has been provided. The voiced/unvoiced detection algorithm used in this analysis is described in the following paragraph.

4.1- Voiced/Unvoiced detection

The voiced/unvoiced detection algorithm is based on the calculation of full band energy, band limited energy, autocorrelation function and zero crossing rate. The signals, that are sampled at 16 kHz (16 bit), are windowed every 5 ms. Hamming windows, 512 samples long, have been used. A first voiced/unvoiced discrimination is made calculating the frame log energy. If this is below a restrictive threshold it is marked as unvoiced and its autocorrelation is not calculated. In this way all pauses are skipped. Then the Short Time Fourier Transform and subsequently the power spectrum are calculated. Actually not all the frequencies of the power spectrum are equally considered but they are weighted in order to give more emphasis to frequencies below 2 kHz (a sort of low-pass filtering) [4]. At this point band limited energy is calculated in the frequency domain and the second energy discrimination is achieved; in this case the energy threshold will be set in order to skip unvoiced fricatives.

On the contrary, voiced fricatives have significant energy in this band of frequencies. In fact, eliminating the high frequency noise, the periodic components in the 0-2 kHz range give a significant contribution to this kind of analysis. The autocorrelation function is obtained calculating the inverse Fourier Transform of the power spectrum. Since the filtered power spectrum has been used, a modified autocorrelation function is obtained and a peak analysis can be made on this function [5]. In practice the highest peaks are selected and their positions, their distances and their periodicity are considered together with the band limited energy, to decide if the frame is voiced or not (figure 2). A set of initial estimates of voiced intervals, for each signal, is obtained and subsequently a refinement procedure is applied.

First of all unvoiced gaps that are one or two frames long are eliminated. In a second time a refinement boundary procedure based on the zero crossing rate calculation is applied. Analysis windows, 5 ms long, are applied forward for the beginning and backward for the ending boundaries of the voiced interval. For each window the zero crossing rate is calculated and if it exceeds a fixed threshold the boundary position is shifted (at most 5 shifts are allowed).

![Figure 2. Voiced/Unvoiced detection algorithm.](image-url)
5. Results and Discussion

As regards the performance of the LE check tool, indicative results are shown in Table 1. They refer to an experiment in which a corpus composed by 1369 utterances pronounced by a male Italian speaker has been used, and in which 24 insertion errors (about 2% of the material) have been simulated by modifying some signals transcriptions. The check process has been performed and the error coefficients described in (3) have been calculated for each sentence. Varying the error coefficient threshold, part of the sentences analyzed has been selected. In particular the number of errors detected at various selections is reported in the table.

<table>
<thead>
<tr>
<th>% Signals selected</th>
<th>Signals selected</th>
<th>% Errors detected</th>
<th>Errors detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>15%</td>
<td>205</td>
<td>100.00%</td>
<td>24</td>
</tr>
<tr>
<td>10%</td>
<td>137</td>
<td>91.67%</td>
<td>22</td>
</tr>
<tr>
<td>5%</td>
<td>68</td>
<td>29.17%</td>
<td>7</td>
</tr>
<tr>
<td>1%</td>
<td>13</td>
<td>16.67%</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Language</th>
<th>Phonemes checked</th>
<th>v/u detection rates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italian</td>
<td>60694</td>
<td>99.11%</td>
</tr>
<tr>
<td>British English</td>
<td>30600</td>
<td>90.31%</td>
</tr>
<tr>
<td>Brazilian</td>
<td>20536</td>
<td>95.81%</td>
</tr>
</tbody>
</table>

Table 1: LE errors detected depending on the percentages of the selected items.

The VU check tool has been tested using three databases of the Italian, British English and Brazilian languages. The detected voiced intervals have been compared with the phonetic transcriptions and segmentations that in this case can be considered consistent, since these languages are at an advanced stage of development. In these tests we assume that a phoneme has to be considered voiced if at least 50% of its duration is voiced. The results are shown in Table 2; the number of checked phonemes is also reported.

<table>
<thead>
<tr>
<th>Phonetic context</th>
<th>Number of occurrences</th>
<th>Number of v/u errors</th>
<th>v/u error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>s d</td>
<td>41</td>
<td>25</td>
<td>61%</td>
</tr>
<tr>
<td>s Dg</td>
<td>50</td>
<td>39</td>
<td>78%</td>
</tr>
<tr>
<td>s v</td>
<td>15</td>
<td>14</td>
<td>93%</td>
</tr>
</tbody>
</table>

Table 2: Voiced/Unvoiced detection rates.

The lower performance of the English and Brazilian languages has been examined through a more precise analysis and it has been discovered that some phonemes have voiced or unvoiced behavior depending on the context in which they are. The following tables report some examples.

<table>
<thead>
<tr>
<th>Phonetic context</th>
<th>Number of occurrences</th>
<th>Number of v/u errors</th>
<th>v/u error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>- b</td>
<td>86</td>
<td>43</td>
<td>50%</td>
</tr>
<tr>
<td>s b</td>
<td>11</td>
<td>11</td>
<td>100%</td>
</tr>
<tr>
<td>- d</td>
<td>16</td>
<td>14</td>
<td>88%</td>
</tr>
<tr>
<td>s d</td>
<td>29</td>
<td>23</td>
<td>79%</td>
</tr>
<tr>
<td>- g</td>
<td>5</td>
<td>5</td>
<td>100%</td>
</tr>
<tr>
<td>s g</td>
<td>11</td>
<td>8</td>
<td>73%</td>
</tr>
</tbody>
</table>

Table 3: examples of Voiced/Unvoiced mismatch in British English.

Table 3 refers to the /b/, /d/ and /g/ phonemes in the British English. When they are at the beginning of a phrase and when they are preceded by an unvoiced consonant they tend to be unvoiced.

Table 4 shows some errors that have been found in the Brazilian data and they refer to the /s/ phoneme that, when followed by a voiced consonant, tends to become vocalized.

<table>
<thead>
<tr>
<th>Phonetic context</th>
<th>Number of occurrences</th>
<th>Number of v/u errors</th>
<th>v/u error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>s d</td>
<td>41</td>
<td>25</td>
<td>61%</td>
</tr>
<tr>
<td>s Dg</td>
<td>50</td>
<td>39</td>
<td>78%</td>
</tr>
<tr>
<td>s v</td>
<td>15</td>
<td>14</td>
<td>93%</td>
</tr>
</tbody>
</table>

Table 4: Examples of Voiced/Unvoiced mismatch in the Brazilian data.

These results show that absolute generalizations are not possible, however some useful indications can be deducted from this kind of analysis in order to implement phonetic transcription rules.

In any case the main goal of this check analysis is to detect discrepancies in the database transcriptions that could have bad effects at synthesis level.

6. Conclusions

A tool to detect some errors in the speech databases used in synthesis systems has been described in this paper. In particular pauses (and in general low energy intervals) and voicing coherence have been checked. This framework has proven to be useful in reducing the manual effort spent in the correction of the databases.

More features will be investigated in order to improve this analysis. In particular, prosodic features will be examined, as they are an important constraint in the unit selection criteria of corpus based text-to-speech systems.

7. References


