Speech Synthesis enhancement in noisy environments

Davide Bonardo, Enrico Zovato
Loquendo S.p.A. - Voice Technologies – Turin, Italy
{davide.bonardo,enrico.zovato}@loquendo.com

Abstract
This paper reports recent activities made to improve the intelligibility of synthesized speech in noisy environments. Nowadays Text-To-Speech technologies (TTS) are used in many embedded devices like mobile phones, PDAs, car navigation systems, etc. This means that speech can be produced in different types of environments where background noise can significantly degrade the perception of the synthetic message and consequently its intelligibility.

The features discussed in this paper are being developed and assessed inside the EU funded SHARE project whose goal is to develop a multimodal communication system supporting rescue operations and disaster management.

1. Introduction
The SHARE project [1] is an IST EU-project started in November 2004 and lasting 36 months. Its goal is to develop an advanced mobile service that will provide critical multimodal communication support for emergency teams during rescue operations. At present, emergency forces use half-duplex channel walkie-talkie technology and are restricted to simple push-to-talk voice communication. After the introduction of the new Push-to-Share service, rescue operations will benefit from sophisticated multimodal interaction and on-line, on-site access to data services providing up-to-date operation status information, as well as details concerning aspects of the emergency, such as location and environment.

The Push-To-Share service developed by SHARE is intended to combine advanced, innovative technologies to allow mobile workers to communicate naturally and bidirectionally, and to share structured multimodal information resources, including audio, video, text, graphic and location information. The Push-To-Share service will make available a full end-to-end solution that will provide rescue teams with a mobile work environment that is intelligent, robust and easy-to-use and which will be evaluated by the Fire Department of Dortmund (Germany).

In this context the TTS (Text-to-Speech) technology is exploited. In fact, if written information needs to be made accessible when hands and eyes are busy, or if standard messages or alerts need to be automatically provided, a natural solution is to rely on systems that are able to read aloud written text with an intelligible voice. Consequently, among the objectives of the SHARE Project is the integration of a high-quality TTS system, able to convert any written (German or English) text into speech, and its improvement and adaptation to the noisy conditions typical of the intended application.

In the next paragraphs we will briefly illustrate the available speech synthesis technology and introduce the features that we have designed to enhance the speech synthesis signals in noisy conditions like a rescue scenario but also like in everyday situations such as listening to news outdoors through a mobile phone or listening to driving directions provided by an in-car navigation system. In section 3 we will describe the feature introduced in the signal processing module of the system and we also report some preliminary analysis and tests results. Section 4 deals with the user controls and shows how they can be effectively used to improve speech perception. The article closes with conclusions and some considerations.

2. TTS technology in noisy environments
The TTS system used in the SHARE project is the one developed at Loquendo and is a multi-voice and multi-language system employing a unit-selection concatenative synthesis technique. Speech units are selected run-time from labeled acoustic databases providing phonetic and prosodic coverage of the intended language. The software yields synthetic human sounding voices whose quality greatly depends on the accuracy of speech databases and on the language dependent knowledge (grammatical, phonetic and prosodic) [2].

The Loquendo TTS technology is integrated in the client side of the SHARE system, in PCs, laptops and tablet PCs; in the future it could be integrated also in PDA devices.

In those environments where the background noise is not negligible, the intelligibility of TTS deteriorates, even though the users/listeners might be equipped with earphones. Hence the need arises to enhance the overall speech intelligibility in order to obtain maximum effectiveness in operative environments. As intelligibility depends both on pronunciation accuracy and on the acoustic quality of the synthetic signal, during the past two years of the Project much effort was devoted to improving linguistic knowledge-bases and obtaining more fluent and more audible speech. Moreover, in order to improve these aspects of the speech synthesis we have mainly worked on signal processing and on the system user controls.

In the first case, the proposed solution is an algorithm to improve the acoustic rendering. It consists of a nonlinear
elaboration of the dynamic range (see section 3) and it involves the signal processing part of the synthesis system. It could be used off-line, preprocessing the TTS speech databases, or be used run-time, activating this processing with the appropriate control.

Regarding user controls, in order to improve the intelligibility in specific domains and applications, the Loquendo TTS offers the possibility of controlling a large number of parameters, both from the system where the engine is integrated through the TTS API (Application Programming Interface) and directly from the text to be synthesized (by introducing control tags into the text to synchronize the control with the reading of the text). One important aspect is that it is possible to combine multiple commands by defining various sorts of macro settings which are able to control the reading (see section 4).

3. Dynamic range controller

Dynamic range controllers like compressors and limiters are used in broadcasting applications to avoid signals exceeding certain peak levels thus provoking unrecoverable damages to the electronic devices [3].

These systems are capable of increasing the perceived loudness while keeping the original intensity range. One of the most important side effects of these controllers is the enhancement of certain sound components and for this reason they are also widely used in music applications. Applied to speech signals they enrich some sounds making them more easily perceivable. Specifically, compressors both smooth the dynamic range of the input signal and increase the apparent loudness and sustain.

3.1. Implementation

The dynamic range controller here introduced, has been designed and implemented to treat signals of TTS databases. The main goal was to synthesize speech so that it is more easily perceivable in noisy environments without degrading its naturalness and quality. For this reason this feature can be used also in quiet environments.

The controller is composed of a compressor with compensation gain: this means that firstly the sound input may be in some cases attenuated and then the whole waveform is amplified in order to re-establish the original peak levels (or higher if necessary). In the processing chain, in fact, a gain value is associated with each input sample in order to attenuate those samples that exceed a fixed threshold. This attenuation depends on the input sample values according to a function, also called static curve, that yields output target level given the input level.

The controller is composed of three modules [4]:

1. the input level detector
2. the input to output level converter
3. the adaptive gain controller.

The first stage is used to measure the input intensity level, whose value is input into the static curve to calculate the target output level. Finally, the adaptive gain controller yields the actual gain value and consequently the new sample output value.

The dynamic range controller scheme is shown in Figure 1. The level detector is a crucial part of the system and its task is to estimate the input intensity level. To this end many methods can be used but the most important consideration is to avoid instantaneous calculation of this parameter because in this case we would have continuous gain oscillations that could cause a sort of “pumping” effect. For this reason most of the algorithms that detect the signal intensity level yield averaged values of the input intensity, and in this sense they behave similarly to the human ear.

We have used a peak level detector that is configured by setting the two parameters $T_a$, the attack time, and $T_r$, the release time. The first parameter describes how fast the detector follows the input when an increase of the input signal level above a certain threshold occurs, while the second one specifies how fast the detector follows the input signal when this falls below a threshold. In our application $T_a$ is set to 0.1 ms and $T_r$ to 300 ms. The general structure of this stage is that of a dynamic first order digital filter whose coefficient is defined by the combination of $T_a$ and $T_r$ [5]. This module has an intrinsic delay of one sample.

The second stage of the controller is the gain function that simply converts the estimated input level to the new output level and indirectly yields the gain coefficient. The design of this function (or static curve) has to follow certain criteria like, for example, the control of the harmonic distortion as well as the preservation of the naturalness of the input sounds.

The static curves that we have implemented, belong to the family of the so called “soft-knee” curves. They are obtained by concatenating straight lines that smoothly change their slope for input values greater than an established value. The initial slope of the curves is almost 1.0, and this means that output values are equal to input values, while when input values exceed about −20dB, output values are increasingly reduced. The final curves are obtained through the cubic interpolation of the original straight lines (see figure 2). In this way the gain values do not abruptly change near the threshold point.

We have designed five curves that gradually increase their compression degree. The curve “sc1” is the one that introduces less compression: in fact at 0dB level (maximum amplitude) there is an attenuation of almost 7 dB. The last curve, named “sc5”, introduces the highest compression degree and should be used in very noisy conditions. In this case the 0dB level is attenuated by 13 dB. Lower intensity values will be gradually attenuated by less than these maximum values and in any case, for all these curves, the compensation gains, that are applied in the last stage, restore the original peak levels so that the dynamic range is the same as the original one, but the processed sound is louder.

The controller is composed of three modules [4]:

1. the input level detector
2. the input to output level converter
3. the adaptive gain controller.
The value obtained through the static curve is the input of the final stage, that is an adaptive gain controller, whose task is to smoothly increase or decrease the gain factor in order to avoid rapid variation in the output signal level. Also this module can be seen as a dynamic first order digital filter whose behaviour is determined by the combination of the two time constants $T_a$, the attack time, and $T_r$, the release time. The meaning of these time constants is the same as that described in the envelope detector module. In our applications these parameters have been set respectively to 1 ms and 250 ms.

3.2. Experiments and results

The main advantage of this kind of processing is increased control over the dynamic range of the speech signal, avoiding saturations and consequently distortions. This is not however obtained at the expense of the average intensity because energy is redistributed, reducing its variability.

The perceived major gain is also confirmed by numerical analysis in which two databases, containing more than two hours of speech recordings, namely from a female German speaker and an English female speaker, were processed with the dynamic range controller (DRC). The overall gain was set in order to maintain peak levels almost equal to the original samples. Data show that the average intensity level increases up to 3.5 – 4.0 dB, depending on the selected static curve (see Figure 3). The control of the maximum peak level is really important as we know the maximum increase in the general volume that we can set without saturating the signal. This means that the DRC can be used together with the TTS controls, and particularly the volume control, in an effective manner in order to improve the perception of the synthesized speech, even in critical conditions.

We have also analysed the spectral effects of this type of elaboration on many voices and the result is that no significant harmonic distortion is introduced. This is mainly due to the smooth static curve and the adaptive stages that the dynamic range controller is composed of. The main result is, on the contrary, the enhancement of some harmonics in voiced sounds but this is perceived as a slightly richer signal.

We have carried out some perceptual tests to evaluate the level of TTS signal degradation due to the introduction of the dynamic range controller. Exploiting binary comparisons, some volunteers had to judge the acoustical quality of two versions of synthesized sentences. The first version, with respect to the second one, was processed with the dynamic range controller, by setting up the sc1 compression degree. The result of this test was that in most of the evaluations, listeners didn’t notice any difference between the two versions, and the percentage of the preferred sentences without DRC is only slightly greater than the percentage of the better evaluated versions with DRC (see Table 1).

<table>
<thead>
<tr>
<th></th>
<th>DRC = NO-DRC</th>
<th>Better DRC</th>
<th>Better NO-DRC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>74.3%</td>
<td>11.4 %</td>
<td>14.3 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>NO-DRC</th>
<th>DRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st syllable</td>
<td>53.8%</td>
<td>58.1%</td>
</tr>
<tr>
<td>2nd syllable</td>
<td>66.2%</td>
<td>70.7%</td>
</tr>
<tr>
<td>3rd syllable</td>
<td>40.0%</td>
<td>47.2%</td>
</tr>
</tbody>
</table>

Table 1: Results of the binary evaluation test on the synthesized speech acoustic quality.

We have also carried out a test to evaluate the intelligibility of the elaborated signals. To this end we have synthesized two sets of non-sense words composed of three syllables. The words were the same in the two sets but in the second case the DRC processing was applied. Then we have added synthetic white noise to all these samples in order to obtain a signal to noise ratio of about 6dB. We have conducted a test in which the same volunteers were asked to listen to these words only once and to transcribe them. The result is that even in case of minimum compression (a DRC with the curve sc1) there is an improvement in the intelligibility of these words (see Table 2).

Table 2: Syllable recognition rates of non-sense words, synthesised without and with DRC (sc1).
4. User controls

Interaction with the Loquendo TTS engine is possible using a large set of generic APIs and control tags essential to achieve the optimal results by controlling some aspects of the TTS reading, such as the language in which the text will be pronounced, the voice, its speaking rate, its loudness, the interpretation of digits, the stress prominence of a word or its pronunciation, and, for specific applications, the activation of sound effects like balance and reverb in the output speech.

In general we can speak of User Controls that allow the adaptation of speech quality and pronunciation to the application conditions.

The User Controls have two types of interactions: asynchronous or synchronous.

The general asynchronous User Controls are the APIs, useful when it is possible or necessary to interact with the speech technology through a user interface, generally using buttons, slider bars and menus. The synchronous User Controls are the control tags, commands inserted in the sentences at the exact point where the control must take effect.

The same control may be set whether asynchronously or synchronously; for example it is possible to increase the volume level with an application button (if we want to listen louder) or by putting the command at a special point in the text where the dialogue requires a speech loudness increase.

When the environmental conditions become critical in terms of increase in noise, it could be necessary to use many of these controls in order to improve the intelligibility. The first step is to obtain the best pronunciation quality: this is made possible by building and activating lexicons to correctly expand domain specific text, such as abbreviations, numbers and street names. For this purpose, during the SHARE activities the lexicon containing the phonetic transcription of the street names of the city of Dortmund has been created with the SAMPA [6] phonetic alphabet.

Other parameters cannot be statically defined: the level and type of noise for example, because they depend on the real-life situation and on the type of environment where the persons and the equipment involved in the rescue operation are. It is important to have the possibility of interacting with the TTS at any time and to have a faster response from the system; this ensures that the message, the alert or the information will be received on time.

With the study of the environment where the TTS will be used, it is possible to define domain-dependent simplified control settings. In practice a single command changes many parameters at the same time and activates specific algorithms and plug-ins. These can be considered as Environment Control Objects (ECO), i.e. collections of settings associated with specific usage conditions.

In noisy environments the volume increase could not be enough to make a message understandable, for this reason the setting of more features is included in the ECO objects.

In fact, with this macro settings command, beyond the volume increase it is also possible to activate the dynamic range control with the appropriate parameter, to slow down the speech rate, and to change other prosodic parameters useful to increase TTS intelligibility.

In the following example, a command message is given using the female British English voice Kate, in a noisy environment: it is necessary to set the volume to the maximum value, select the appropriate dynamic range preset, and pronounce the sentence slower than the normal speech rate:

```
ECO \voice=Kate \volume=100
\dynamicrange=5 \speed=45
"Switch off the electric power immediately"
```

The collection of control tags sequence in the previous example represents an ECO object. The idea is that a user could define the ECO objects useful for its applications or services and apply them when necessary.

5. Conclusions

The goal of the project is to improve the intelligibility of speech synthesis in noisy environments. To this end, we have shown that it is possible to effectively enhance the output signal just by working on the acoustic stage of the system. At the same time, an effective use and integration of the available user controls permits the configuration of the system according to the application environment, thus yielding the necessary flexibility.

Improvements at algorithmic level will be considered to better simulate what people do when they speak in noisy environments. Particularly, techniques to effectively simulate spectral and pitch changes, without degrading the speech synthesis quality will be investigated. Further experiments will be also carried out in different environments in order to evaluate the effects of the combination of these features.

A further development will be to design an adaptive control setting, by which on the basis of SNR (Signal to Noise Ratio) or other noise indicators the TTS will be adapted online to the external environmental conditions. For this reason either a SNR processing or an ASR engine may be used to determine online the environmental conditions. With this information it is possible to have an automatic control for the TTS that can modify at run time its control settings in order to adapt the intelligibility to external conditions.

6. Acknowledgements

We would like to thank Jobst Löffler from the Fraunhofer IAIS (Institut Intelligente Analyse und Informationssysteme) and the Fire Department of Dortmund (Feuerwehr Dortmund) for providing the audio data used in the first experiments.

7. References