Improving speech synthesis quality by reducing pitch peaks in the source recordings

Luisina Violante, **Pablo Rodríguez Zivic** and **Agustín Gravano**
Departamento de Computación, FCEyN
Universidad de Buenos Aires, Argentina
{lviolante,prodriguez,gravano}@dc.uba.ar

Abstract

We present a method for improving the perceived naturalness of corpus-based speech synthesizers. It consists in removing pronounced pitch peaks in the original recordings, which typically lead to noticeable discontinuities in the synthesized speech. We perceptually evaluated this method using two concatenative and two HMM-based synthesis systems, and found that using it on the source recordings managed to improve the naturalness of the synthesizers and had no effect on their intelligibility.

1 Introduction

By definition, corpus-based speech synthesizers, such as concatenative and HMM-based systems, rely heavily on the quality of the speech corpus used for building the systems. Creating speech corpora for this purpose is expensive and time consuming, so when the synthesized speech obtained is not as good as expected, it may be desirable to modify or correct the corpus rather than record a new one. Common corrections are limited to discarding mispronounced words or noisy units. In this work we describe a simple method for attenuating pronounced pitch peaks, a frequent problem in recordings made by professional speakers, and evaluate it using four different corpus-based systems. Sections 2 and 3 describe the speech synthesis systems and corpus employed in this work. In Section 4 we present the method for reducing pitch peaks. In Section 5 we describe how we evaluated the effect of our method on intelligibility and naturalness of the synthesizers.

2 Synthesis systems

Festival[1] is a general framework for building speech synthesis systems, written in C++ and developed by the Center of Speech Technology Research at the University of Edinburgh (Black et al., 2001). It provides an implementation of concatenative speech synthesis as well as synthesis based on Hidden Markov Models (HMM). In this work we used a Festival module called *Clunits unit selection engine* to build concatenative synthesizers. The unit size is the *phone*, although since a percentage of the previous unit is included in the acoustic distance measure, the unit size is rather "phone plus previous phone", thus similar to a *diphone* (Black and Lenzo, 2007). Additionally, we used a second Festival module called *Clustergen parametric synthesis engine* for building HMM-based speech synthesizers.

MARY TTS[2] is an open-source synthesis platform written in Java, originally jointly developed by the Language Technology Lab at the German Research Center for Artificial Intelligence (DFKI) and the Institute of Phonetics at Saarland University, and currently maintained by DFKI. Like Festival, MARY provides toolkits for building unit selection and HMM-based synthesis voices (Schröder and Trouvain, 2003).

3 Corpus

For building our systems we used the SECYT corpus, created by the Laboratorio de Investigaciones Sensoriales (Universidad de Buenos Aires) for

[1] http://festvox.org/festival
[2] http://mary.dfki.de

Proceedings of NAACL-HLT 2013, pages 502–506,
Atlanta, Georgia, 9–14 June 2013.

studying the prosody of Argentine Spanish (Torres and Gurlekian, 2004). It consists of 741 declarative sentences recorded by a female professional speaker (pitch range: 130-380Hz). On average, sentences are 7 words and 3.9 seconds long. The entire corpus has manual phonetic transcriptions and time alignments, following a version of the *Speech Assessment Methods Phonetic Alphabet* (SAMPA) adapted for Argentine Spanish (Gurlekian et al., 2001).

A priori, this corpus is a very good candidate for building a synthesis system – its 741 sentences are phonetically balanced, the audio quality is excellent, and it has precise time-aligned phonetic transcriptions. We thus built two concatenation systems using this corpus: Festival's diphone-like and MARY's diphone systems. The results were not satisfactory. The new voices presented clearly noticeable discontinuities, both in intensity and pitch, which affected their naturalness – as judged impressionistically by the authors and non-expert colleagues.

In an attempt to attenuate these problems, we leveled the intensity of all recordings to a mean of 72dB using linear interpolation. Specifically, each sound was multiplied by a number such that its new average RMS intensity was 72dB; so that all sentences in the corpus ended up with the same average intensity. After this conversion, we rebuilt the systems. The resulting voices sounded somewhat better, but their most noticeable problem, severe pitch discontinuities, persisted.

Further analysis of the corpus recordings revealed that this issue was likely due to the speaking style employed by the professional speaker. It contains frequent pronounced pitch peaks, a verbal stylistic device acquired by the speaker as part of her professional training. These events produced units with very different pitch levels and slopes, thus leading to the discontinuities mentioned above.

4 Reduction of pitch peaks

We searched for ways to reduce the magnitude of these pitch peaks by manipulating the pitch track of the recordings using the *Time-Domain Pitch-Synchronous OverLap-and-Add* (TD-PSOLA) signal processing technique (Moulines and Charpentier, 1990). We used the implementation of TD-PSOLA included in the Praat toolkit (Boersma and Weenink, 2012).

We tried several formulas for TD-PSOLA and ended up choosing the one that appeared to yield the best results, evaluated perceptually by the authors:

$$f(x) = \begin{cases} (x - T) * s + T & \text{if } x > T \\ x & \text{otherwise.} \end{cases}$$

This formula linearly scales the pitch track by a scaling factor s above a threshold T, and leaves it intact below T. When $0 < s < 1$, the pitch track gets compressed above the threshold. We experimented with several values for the two constants, and selected $T = 200$Hz and $s = 0.4$ as the ones producing the best results. Figure 1 illustrates the pitch peak reduction method. The black solid line corresponds to

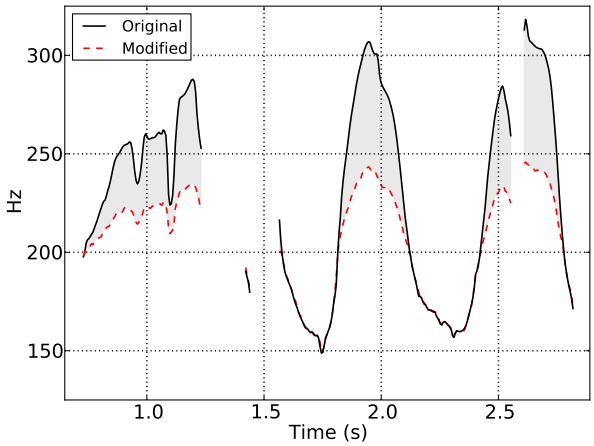

Figure 1: Reduction of pitch peaks. The original pitch track (in black) is scaled down 40% above 200Hz.

the pitch track of the original audio; the red dotted line, to the pitch track of the modified audio. Note that the modified pitch track is scaled down above 200Hz, but identical to the original below it.

5 Evaluation of the method

Next we proceeded to evaluate the effect on synthesizer quality of reducing pitch peaks in the training corpus. For this purpose we prepared two versions of the SECYT corpus – with and without applying our pitch-peak reduction technique. We refer to these two as the *original* and *modified* recordings, respectively. In both cases, the intensity level of all audios was first leveled to a mean of 72dB using linear interpolation, to compensate for differences across recordings.

Subsequently, we built 8 speech synthesizers, consisting in all combinations of: Festival and MARY frameworks, concatenative and HMM-based synthesis, and original and modified recordings. We refer to these systems using the following notation: {fest, mary}_{conc, hmm}_{orig, mod}; e.g., mary_conc_mod is a concatenative system built using the MARY framework with the modified corpus.

We evaluated these systems along two dimensions: intelligibility and naturalness. Our goal was to compare four system pairs: systems built using the original recordings vs. those built using the modified recordings. The null hypothesis was that there was no difference between 'orig' and 'mod' systems; and the alternative hypothesis was that 'mod' systems were better than 'orig' ones.

5.1 Intelligibility

To evaluate intelligibility we used the *Semantically Unpredictable Sentences* (SUS) method (Nye and Gaitenby, 1974), which consists in asking participants to listen to and transcribe sentences with correct syntax but no semantic sense, for later measuring and comparing the number of transcription errors. We used a set of 50 such sentences, each 6-10 words long, created by Gurlekian et al. (2012) for evaluating Spanish speech synthesizers. A sample sentence is, *El viento dulce armó un libro de panqueques* (The sweet wind made a book of pancakes).

For each participant, 40 sentences were selected at random and synthesized with the 8 systems (5 sentences per system, with no repetitions). Participants were given the following instructions,

> *La primera tarea consiste en escuchar varios audios, y transcribir para cada audio la oración que escuches. Prestá atención, porque podés escuchar cada audio una sola vez.*
> (The first task consists in listening to several audios, and transcribing for each audio the sentence you hear. Pay attention, because you can only listen to each audio once.)

5.2 Naturalness

To evaluate naturalness we used the *Mean Opinion Score* (MOS) method, in which participants are asked to rate the overall quality of synthesized speech on a 10-point scale (Viswanathan and Viswanathan, 2005).

We used a set of 20 sentences, each 5-20 words long, created by Gurlekian et al. (2012), plus 20 additional sentences created for this study. A sample sentence is, *El sector de informática es el nuevo generador de empleo del país* (The information technology sector is the country's new job creator).

Again, for each participant, 40 sentences were selected at random and synthesized with the 8 systems (5 sentences per system). Participants were given the following instructions,

> *La segunda (y última) tarea consiste en escuchar otros audios, y puntuar la naturalidad de cada uno. Usar una escala de 1 a 10, donde 1 significa "no suena natural en lo absoluto" y 10 significa "suena completamente natural". En este caso, podés escuchar cada audio una o más veces.*
> (The second (and last) task consists in listening to other audios, and score the naturalness of each. Use a scale from 1 to 10, where 1 means "it does not sound natural at all" and 10 means "it sounds completely natural". In this case, you may listen to each audio one or more times.)

5.3 Results

SUS and MOS tests were administered on a computer interface in a silent laboratory using regular headphones. 14 graduate and undergraduate students (11 male, 3 female; mean age: 27.6) completed both tests – first SUS, followed by MOS.

The transcriptions of the SUS tests were manually corrected for obvious typos and spelling errors that did not form a valid Spanish word. Suspected typos and spelling errors that formed a valid word were not corrected. For example, *peliculas* was corrected to *películas*, and *precion* to *presión*; but *canto* was not corrected to *cantó*, since it is a valid word. Subsequently, we computed the Levenshtein distance between each transcription and the corresponding sentence. Figure 2 shows the distribution of Levenshtein distances for each of our eight systems. We observe that all systems had a low error count, with a median of 0 or 1 errors per sentence. Two-tail Wilcoxon signed-rank tests revealed no significant differences between the systems built with the original and modified recordings ($p = 0.70$ for fest_conc, $p = 0.40$ for fest_hmm, $p = 0.69$ for mary_conc, $p = 0.40$ for mary_hmm, and $p = 0.41$ for all systems together). These results indicate that the intelligibility of all four system types was not affected by the

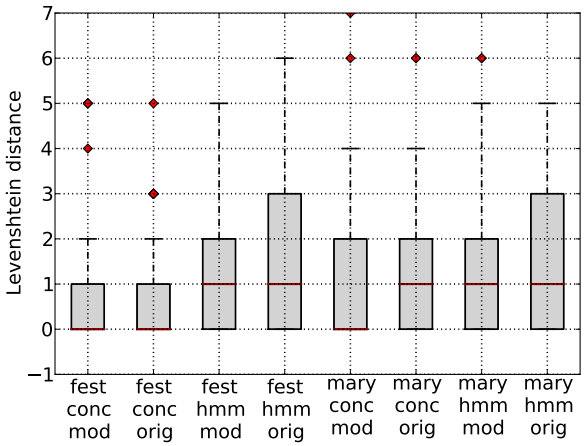

Figure 2: Intelligibility (SUS) results.

modifications performed on the corpus for reducing pitch peaks.

To account for the different interpretations of the 10-point scale, we normalized all MOS test scores by participant using z-scores.[3] Figure 3 shows the distribution of values for each system.

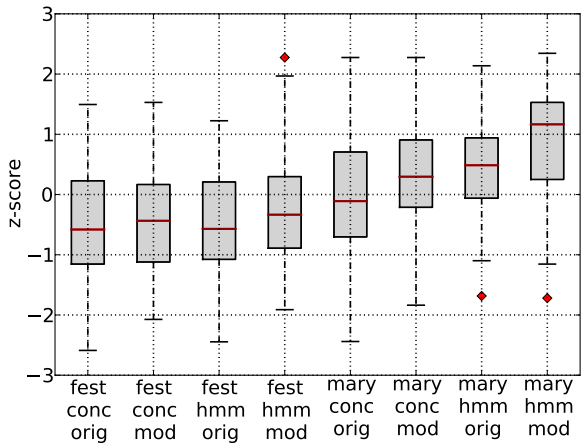

Figure 3: Naturalness (MOS) results.

We performed a series of Wilcoxon signed-rank tests to assess the statistical significance of the observed differences. The null hypothesis was that there was no difference between 'orig' and 'mod' systems; and the alternative hypothesis was that 'mod' systems were perceived as more natural than 'orig' ones. Table 5.3 summarizes these results.

For mary_conc and mary_hmm (concatenative and HMM-based systems built using the MARY framework) the perceived naturalness was significantly higher for systems built using the modified recordings (i.e., after reducing pitch peaks) than for systems built with the original recordings. For fest_conc (concatenative system built with Festival) we found no evidence of such differences. Finally, for fest_hmm (Festival HMM-based) the difference approaches significance at 0.126.

[3] $z = (x - \overline{x})/s$, where $\overline{x}$ and s are estimates of the participant's mean and standard deviation, respectively.

	W	p-value
fest_conc	2485	0.559
fest_hmm	2175	0.126
mary_conc	1933	0.016
mary_hmm	1680.5	0.001
All systems	34064.5	0.004

Table 1: Results of Wilcoxon tests comparing systems using the original and modified audios.

6 Conclusions

In this paper we presented a method for improving the perceived naturalness of corpus-based speech synthesizers. It consists in removing pronounced pitch peaks in the original recordings, which typically produce discontinuities in the synthesized speech. We evaluated this method using two common technologies (concatenative and HMM-based synthesis) and two different implementations (Festival and MARY), aiming at a good coverage of state-of-the-art speech synthesizers, and obtained clear results. First, its utilization on the source recordings had no effect (negative or positive) on the intelligibility of any of the systems. Second, the naturalness of the concatenative and HMM-based systems built with the MARY framework improved significantly; the HMM-based system built with Festival showed an improved naturalness at a level approaching significance; and the Festival concatenative system showed no improvement. In summary, the presented method did not harm the intelligibility of the systems, and in some cases managed to improve their naturalness. Therefore, since the impact of the proposed modifications on all four systems was positive to neutral, developers may find this methodology beneficial.

Acknowledgments

This work was funded in part by CONICET, ANPCYT PICT 2009-0026, and UBACYT 20020090300087. The authors thank Jorge A. Gurlekian, Humberto M. Torres and Christian G. Cossio-Mercado from LIS (INIGEM, CONICET-UBA) for kindly sharing the SECYT corpus and other materials for the present study, as well as for valuable suggestions and comments.

References

Alan W. Black and Kevin A. Lenzo. 2007. *Building Synthetic Voices*. Language Technologies Institute, Carnegie Mellon University, http://festvox.org/bsv.

A. Black, P. Taylor, R. Caley, R. Clark, K. Richmond, S. King, V. Strom, and H. Zen. 2001. The festival speech synthesis system.

Paul Boersma and David Weenink. 2012. Praat: doing phonetics by computer. *http://www.praat.org/*.

J. Gurlekian, L. Colantoni, and H. Torres. 2001. El alfabeto fonético SAMPA y el diseño de corpora fonéticamente balanceados. *Fonoaudiológica*, 47:58–69.

J. A. Gurlekian, C. Cossio-Mercado, H. Torres, and M. E. Vaccari. 2012. Subjective evaluation of a high quality text-to-speech system for Argentine Spanish. In *Proceedings of Iberspeech*, Madrid, Spain.

E. Moulines and F. Charpentier. 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication*, 9(5):453–467.

P. W. Nye and J. H. Gaitenby. 1974. The intelligibility of synthetic monosyllabic words in short, syntactically normal sentences. *Haskins Laboratories Status Report on Speech Research*, 37(38):169–190.

M. Schröder and J. Trouvain. 2003. The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6(4):365–377.

H. M. Torres and J. A. Gurlekian. 2004. Automatic determination of phrase breaks for Argentine Spanish. In *Speech Prosody 2004, International Conference*.

Mahesh Viswanathan and Madhubalan Viswanathan. 2005. Measuring speech quality for text-to-speech systems: Development and assessment of a modified mean opinion score (MOS) scale. *Computer Speech & Language*, 19(1):55–83.