Foreign Accents in Synthetic Speech: Development and Evaluation

Laura Mayfield Tomokiyo, Alan W Black, Kevin A. Lenzo

Cepstral, LLC
Pittsburgh, PA 15206 USA
laura@cepstral.com

10.21437/Interspeech.2005-519

Abstract

This paper addresses the generation and evaluation of foreign-accented speech in concatenative text-to-speech (TTS) synthesis. We describe three possible methods of building a Spanish-accented English voice, and evaluate and compare them with respect to preference, intelligibility, and smoothness. Effects of speaking rate and content are also examined.

It is found that although using an unmodified Spanish voice to read English text is possible, the result is not highly intelligible. With some modifications to the linguistic model, a relatively high level of comprehensibility and smoothness can be achieved, not differing widely from ratings given to a native voice at a comparable stage of development. Listeners in perceptual experiments were very consistent in their preference rankings of the three voices, showing that differences in voice-building method are both detectable and contribute to synthesis quality.

1. Introduction

In a multilingual context, it may be necessary to synthesize words in a foreign language. The optimal approach to generating the foreign-language speech may be quite different for different applications, however. In order to read non-English[1] names in an English context, it may be sufficient to pronounce the name as an educated speaker would, specifically by identifying the name as foreign and applying a specialized grapheme-to-phoneme model to get a probable pronunciation using the English voice [5].Word sequences in different languages can be interleaved, as in e-mail between multilingual friends:

> *Otanjoubi omedetou*! I'm going to just have a *nombiri* golden *wi-ku* – hopefully head out of town for a few days for some nice *onsen* time. *A, sousou...* I sent a *shinkansen omiyage* for *chibi* so you should be getting it any day!

In this situation, we would still probably prefer to use the same English voice seamlessly, but with more detailed information about the Japanese pronunciations and phone mappings between English and Japanese. For multilingual web pages, where the same content is provided in multiple languages, it may be most desirable to identify the language and then switch between native voices. These situations are all quite different, but they all assume that the core of the text to be spoken is in the language native to the synthetic voice. Certainly, the problem of language identification is not trivial. In terms of intelligibility, however, there is a basis for interpreting segments of foreign-accented speech both perceptually, as the listener becomes familiar with the idiosyncrasies of the voice, and semantically, as listeners can draw on context to understand heavily accented segments.

When a synthetic voice is used to speak running text in a foreign language, barriers to intelligibility quickly emerge. In preliminary evaluations of the foreign-accented voices to be described here, it was observed that although proper names were generally understandable, synthesis of unrestricted text had an average intelligibility of nearly zero. This contrasts, incidentally, with the common observation in unit-selection synthesis that single words are *less* acceptable than sentences. Although search algorithms are able to optimize the selection of units for a particular source-target language pair [3], overall intelligibility for complex text of the sort one might encounter in a newsreading task remains extremely low. Foreign-accented voices are becoming increasingly desirable, however, as TTS is used to create a persona: a convincing character for a computer game, for example, or perhaps a voice that sounds like the user's own. Foreign-accented TTS can also add realism to applications like speech-to-speech translation, where a user speaks into a device to have his words translated and spoken in a different language.

This paper reports on the development and evaluation of a Spanish-accented English voice from a native Spanish database. We investigate how intelligibility is affected by two variables: voicebuilding process and speaking rate. We also examine the "forgiveness factor" of accented voices, asking whether synthesis errors are less damaging than in a native voice. Finally, we discuss how intelligibility is affected by the content of the text to be read.

2. Related Work

It has been observed in studies of natural speech that even highly accented speech can be very intelligible [4]. Schmid and Yeni-Komshian propose a number of explanations for this phenomenon, including that listeners subconsciously alter their threshold of acceptability, and that while the non-native speech may be perceived as equally intelligible, listeners are expending more effort in understanding it [8].

The situation for foreign-accented voices in speech synthesis is not as clear. There are two factors contributing to processing cost for the listener: accentedness and synthesis quality. We do, however, have the advantage in synthesis of being able to control the content of the speech; we can eliminate reading errors, infelicitous lexical choice, and unconventional syntax from the equation.

Recent descriptions of the generation and evaluation of foreign-accented synthetic speech have focused on the detection and handling of regions of foreign text in a native-language context. The German language is particularly challenging in this regard, as it is highly inflected and use of English and other foreign-language words is common. The foreign-language words can become part of a mixed-language compound or undergo German declension or conjugation. Pfister and Romsdorfer describe an approach to detecting and morphologically analyzing these regions in a highly accurate way [7].

Once a region of foreign text has been identified, a TTS system has to know how to say it. Badino, Barolo, and Quazza describe a method of first looking up words in the foreign region in a pronunciation model for that language, and then mapping the phonemes onto the closest matches in the synthesis language as determined by a articulatory feature based similarity function [1].

[1] All examples are meant to apply to any language; English is used here for simplicity of reference.

Adaptation in the prosodic space can improve the quality of accented speech generated via phoneme mapping. Campbell [3] optimizes the selection of units by synthesizing the target text in both the output language and the language native to the text, and select the units for foreign-accented synthesis that best match the acoustic characteristics of the native synthesis. The requirement that two synthesizers run in parallel may be prohibitive in some contexts, however.

3. Building an accented voice

Approaches to creating an accented unit-selection voice range from simply using (in this case) a Spanish voice and linguistic model to interpret and speak English text to recording an entire database from an accented speaker. We have held the latter for further experimentation, and evaluate three methods of developing a Spanish-accented voice from Spanish audio data.

3.1. The Swift speech synthesis engine

All experiments described in this paper use Cepstral's Swift™unit-selection synthesis engine.

There are three independent components in the Swift TTS system: the synthesis engine, the voice, and the general linguistic model. The accented voices described here have undergone modifications in the third component, the linguistic model. The lexicon, which specifies the pronunciations of known words, is part of the linguistic model. The linguistic model also controls the token-to-word interpretation (how the text, possibly containing numerals, acronyms, symbols, abbreviations, etc. is converted to lexical words), the grapheme-to-phoneme model (predicting the pronunciation of unknown words), the part-of-speech tagger, the post-lexical model, the prosodic model, and other related models.

3.2. Juan: a Spanish voice speaks English

The baseline voice, *Juan*, is a mature Spanish voice, with no changes to the linguistic model at all. English words are interpreted using the Spanish token-to-word and grapheme-to-phoneme model, resulting in such unacceptable expansions as

I will check in on May 14th
`/i#uil#čekk#in#on#maj#katorseteače/`

The English pronoun *I* is rendered as /i/, and the number sequence 14th is interpreted as the Spanish *catorce* concatenated with the letter sequence *t-h*. A poor model, but surprisingly intelligible for many words.

3.3. Manuel: English linguistic model with a Spanish voice

The *Manuel* voice is a great leap forward from Juan. Manuel uses token-to-word and grapheme-to-phoneme models trained on English data, with context-independent phone mappings derived from articulatory-feature-space distance measures. Mappings are listed in Table 1. One limitation of this particular procedure as implemented in our framework is that only single-phoneme mappings are possible; English /er/ might be better mapped to Spanish /e r/ but the system is forced to generate only one phone, namely /r/. Phoneme mappings are applied runtime; that is, when the English linguistic model (lexicon and/or grapheme-to-phoneme model) requests a particular phoneme, the mapping table is consulted and a corresponding Spanish unit is generated.

Because this is treated as an English voice, the normal English prosodic model is imposed on it; the search will favor unit sequences that are prosodically typical of English.

EN	ES	EN	ES	EN	ES
h	x	aa	a	ih	i
jh	ch	ae	a	ow	o
ng	n	ah	a	oy	o
sh	s	ao	o	uh	u
th	t	ay	a	uw	u
v	b	dh	d		
w	xu	eh	e		
z	s	er	r		
zh	ch	ey	e		

Table 1: Phone mappings for the Manuel voice.

ES	HS	EN	ES	HS	EN
a	a	a, ae	ch	ch	ch, jh, zh
e	ey	ey, eh	d	d	d, dh
i	i	i, ih	n	n	n, ng
o	ow	ow, oa, ao	rr, r	r	r
u	uw	uw, uh	s	s	s, sh, z
e r	ey r	er	t	t	t, th
a xi	a j	ay	v	b	b, v
o xi	ow j	oy	x	h	h
xi, j	j	j			
xu	w	w			

Table 2: Mappings from English and Spanish to the simplified **HS** (Hispanic) phone set in the Antonio voice. /xi/ and /xu/ are the weak vowels found in diphthongs such as [aĭ] ("hay") and [ŭa] ("cuanto").

3.4. Antonio: English linguistic model trained with Spanish data

Some of the limitations of Manuel are addressed in the *Antonio* voice. In Antonio, the linguistic model is adapted to the foreign-accented condition. Phoneme mappings are encoded directly in the lexicon, permitting both multiple-to-single and single-to-multiple mappings. The grapheme-to-phoneme model is trained on this lexicon, meaning that unknown words are synthesized according to their likely pronunciation by a Spanish-accented voice. Context dependencies, which specify phone sequences that trigger allophonic variation and are best selected from similar contexts, are also trained on the accented lexicon. Although the list of potential dependencies itself is not adapted, the runtime application is optimized to reflect the sequences that actually appear in the Spanish-language database.

The linguistic model for the Antonio voice uses a simplified phoneset, shown in Table 2. As in the Manuel voice, the normal English prosodic model is imposed.

There was no manual tuning of the phonetic labels in either the Manuel or Antonio voices to optimize synthesis on the new linguistic models. It is expected that the voices would see a measurable improvement after tuning.

4. Evaluation

The three voices described above were evaluated for overall acceptability, preference, intelligibility at different speaking rates, and smoothness. A fourth voice, a native English voice that had undergone a comparable degree of tuning, was included in acceptability and smoothness tests to evaluate the forgiveness factor, or *intelligibility benefit*[2] of an accented voice.

4.1. Methodology

4.1.1. Evaluation participants

Listening tests were carried out by six adult native speakers of English. All listeners were phonetically aware and had been exposed to accented English before.

4.1.2. Test data

The sentences used in the listening tests were derived from English as a Second Language (ESL) teaching material. Sentences ranged from five to fifteen words in length. Example sentences are:

Can you tell me how to find a bookstore around here?
I was born and raised in Colombia.
I would like to have the chicken soup, please.

Initially, it was expected that evaluation sentences would be the same as for English voices, with sentences in the news domain and no attempt to control grammatical complexity. Preliminary experiments showed that the voices were not suited for complex sentences, although it was observed that intelligibility did improve when the speaking rate was artificially reduced to reflect the naturally slower pace of non-native speech [6].

Two types of simplified text were then tested, and a noticeable content effect was observed. Sentences of similar length and difficulty of vocabulary were taken from the children's news publication *Time for Kids* and ESL sample dialogues. While the ability of a listener to transcribe the synthesized sentence without seeing the text (blind intelligibility) was fair for the ESL sentences, it was almost zero for even the simplified news sentences.

Seventy ESL sentences were selected for inclusion in the evaluation. Of those, twenty were used for each series of preference tests (pairwise at normal speed, pairwise at reduced speed, voice-internal speaking rate). The remaining ten were used for the smoothness and acceptability tests.

4.1.3. Testing environment

Evaluations were presented to the listeners using a web-based interface, with audio played through headsets. Three types of tests were conducted.

Preference test Listeners are presented with a pair of synthesized utterances and asked to check the box next to one they prefer.

Acceptability test Listeners were presented with a single synthesized utterances and asked to give Mean Opinion Score (MOS) rating on a 5-point Likert scale, with 1=bad and 5=excellent.

Smoothness test Listeners are presented with a single synthesized utterance and asked to check a box next to each word that has synthesis errors.

4.2. Voice preference test

4.2.1. Normal speed

The three accented voices were tested in round-robin format in voice preference (A/B) tests. That is, for each of the 30 sentences in the relevant test set, speakers were asked to choose between Juan and Manuel, between Manuel and Antonio, and between Antonio and Juan. Listeners were not given specific objective criteria for making the judgment. Results, showing the percentage of the listeners that preferred each voice, are given in Table 3.

It is clear that Juan is the worst voice. In only 3% of cases was Juan preferred to Manuel, and in only 1% of cases was Juan preferred to Antonio.

Antonio emerged as the best voice, being preferred over Juan in 99% of cases and over Manuel in 70% of cases.

4.2.2. Reduced speed

As noted in Section 4.1.2, initial observations had suggested that artificially reducing the speaking rate can increase the intelligibility of the accented voice. To further explore this, we repeated the voice preference experiments with synthesis slowed to 120 words/minute (compared to approximately 150 words/minute for the naturally synthesized voice).

Voice A			Voice B	
Manuel	30%		70%	Antonio
Juan	3		97	Manuel
Juan	1		99	Antonio

Table 3: Percent of judgments for Voice A vs. Voice B.

Voice A			Voice B	
Manuel	40%		60%	Antonio
Juan	5		95	Manuel
Juan	0		100	Antonio

Table 4: Percent of judgments for Voice A vs. Voice B for artificially slowed synthesis.

Results are shown in Table 4. Comparing the values in Table 4 with those in Table 3, we see that the Manuel voice enjoys the greatest benefit from the speaking rate reduction. He goes from being preferred to Antonio in 30% of cases to 40%. Juan also showed a small advantage from slowing.

4.3. Speaking rate preference test

To follow up on observations on speaking rate, direct preference tests were conducted on each voice to compare intelligibility of normal- and reduced-speed voices. Reduced-speed voices were slowed to 120 words/minute (compared to approximately 150 words/minute for the naturally synthesized voice). Results are shown in Figure 1.

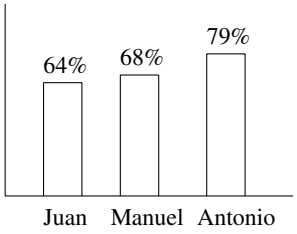

Figure 1: Percentage of judgments favoring normal speed over synthesis artificially slowed to 120 words/minute.

It should be noted that for no test involving a speaking rate change were listeners told that speaking rate had been modified. Although the normal and slowed voices did sound different (all results in this paper clearly support this), the speaking rate was not reduced so dramatically that listeners knew it was a rate reduction, as opposed to a slow-speaking model speaker. The normal-speed and slowed voices were treated as separate voices for the purposes of evaluation.

4.4. Overall acceptability test

Overall acceptability (MOS) tests were conducted for all three voices at both speeds, along with a native speaker at both speeds. Listeners were asked to rate the acceptability of each synthesized utterance on a Likert scale of 1 (bad) to 5 (excellent).

Results are given in Table 5.

4.5. Smoothness test

The goal of the smoothness test is to obtain an objective measure of synthesis quality. For each word in a sentence, listeners

	normal	slow
Juan	2.7	2.3
Manuel	4.2	3.8
Antonio	4.3	4.1
Native	4.7	4.4

Table 5: MOS scores (1-5) for the three accented voices and one native voice at normal and reduced speeds.

are asked to decide whether the word was synthesized clearly and smoothly ("good") or not ("bad"). The overall smoothness score for each sentence is the ratio of good words to total words.

Results are shown in Table 6. Although it had been anticipated that the native voice would be much smoother, all but the Juan voice were judged as relatively smooth. Poorer ratings of (otherwise identical) reduced-speed sentences suggest that synthesis errors are more evident in slower speech.

	normal	slow
Juan	75%	62%
Manuel	92	87
Antonio	90	88
Native	94	90

Table 6: Smoothness scores for the three accented voices and one native voice at normal and reduced speeds.

The Antonio voice loses its advantage over the Manuel voice in this context, although Antonio's MOS scores are slightly higher for the same sentences. One possible explanation is that join errors are more damaging for the higher-quality voice.

5. Discussion

There was a noticeable effect of content in the intelligibility of the sentences. Sentences of comparable length from children's news text were much less intelligible than sentences from the ESL practice dialogues. A possible explanation is that the listener is better prepared to hear accented sentences in the ESL domain than even simplified sentences in the news domain. The news domain sentences did contain informal expressions ("When you think of dinosaurs, what do you picture?") while the ESL sentences were fairly straightforward ("Can you tell me how to get downtown by public transportation?"). The impact of content on intelligibility of accented synthesis is a target for further exploration.

The phoneme mappings described in this paper were completely context-independent. It is likely that by taking phonetic context into consideration a better phonetic match could be achieved. For example, the US English flap, found in words like "bu*tt*er" is phonetically close to the Spanish alveolar [r]. With a context-dependent mapping, we might be able to substitute [r] instead of [t] in this case. Similarly, although there is no /dh/ phoneme in American Spanish, it does occur phonetically as an allophonic variant of /d/ in intervocalic contexts. With more sophisticated context-dependent mapping we could insert a [dh] unit where /dh/ is requested by the English linguistic model.

It is not clear, however, whether such context-dependent phoneme mapping would produce more intelligible synthesis. It is possible that native speakers of English would *expect* a Spanish-accented speaker to pronounce "writer" with a [t], and the violation of this expectation might actually increase the processing cost.

The impact of the prosodic model on foreign-accented synthesis has not been formally evaluated. The prosodic model can affect both intelligibility and perception of accentedness.

It had been hypothesized that artificially slowing the speaking rate would make the synthesis easier to understand, as the native listener would have more time to process the unusual pronunciation. The results presented here suggest that for concatenative synthesis, this is not the case. It does appear that slowing down the synthesis is more damaging for the higher-quality voices. The synthesized sentences were grammatically and lexically very simple, however, and the impact may be different for more difficult text. Artifacts introduced through signal processing in the rate reduction process may also contribute to reduced intelligibility.

Expanding the evaluation to include a Spanish-accented English voice database would also be a natural extension of this work. In recordings of Spanish-accented English, we would have a phoneme context inventory that is much closer to that of native English, and could make better use of contextual modeling.

6. Summary

This study evaluates different methods of generating a Spanish-accented English voice from native Spanish recordings in unit-selection synthesis. Three accented voices were tested: one using a Spanish unit database and Spanish linguistic model; one using a Spanish unit database and unadapted English linguistic model, and one using a Spanish unit database and an English linguistic model adapted to Spanish.

In experiments testing preference, acceptability, and smoothness of the three Spanish-accented English synthetic voices, it was found that the combination of Spanish audio data with an English linguistic model adapted to reflect the phonemes and phoneme combinations found in the Spanish data performed best.

When compared with a native voice for acceptability (MOS) and smoothness tests, both of the accented voices using the English linguistic model performed comparably with the native voice, although the native voice was rated slightly higher.

All experiments were repeated with the speaking rate artificially slowed. Direct comparisons between normal- and reduced-speed synthesized sentences were also conducted. The normal-speed voices were found to be superior in every case.

7. References

[1] Leonardo Badino, Claudia Barolo, and Silvia Quazza. Language independent phoneme mapping for foreign TTS. In *IEEE Workshop on Text-to-speech Synthesis*, 2004.

[2] Tessa Bent and Ann R. Bradlow. The interlanguage speech intelligibility benefit. *Journal of the Acoustical Society of America*, 114(3):1600–1610, 2003.

[3] Nick Campbell. Talking foreign. In *Proc. Eurospeech*, pages 337–340, 2001.

[4] Tracey Derwing and Murray Munro. Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 19:1–16, 1997.

[5] Ariadna Font Llitjos. Improving pronunciation accuracy of proper names with language origin classes. Master's thesis, Carnegie Mellon University, August 2001. CMU-LTI-01-169.

[6] Laura Mayfield Tomokiyo. Linguistic Properties of Non-native Speech. In *Proc. ICASSP*, 2000.

[7] Beat Pfister and Harald Romsdorfer. Mixed-lingual text analysis for polyglot TTS synthesis. In *Proc. Eurospeech*, 2003.

[8] Peggy M. Schmid and Grace H. Yeni-Komshian. The effects of speaker accent and target predictability on perception of mispronunciations. *Journal of Speech, Language, and Hearing Research*, Feb. 1999.