Expanding Phonetic Coverage in Unit Selection Synthesis through Unit Substitution from a Donor Voice

Alistair Conkie and Ann K. Syrdal

AT&T Labs – Research
Florham Park, NJ 07932-0971 U.S.A.
adc,syrdal@research.att.com

Abstract

This paper describes experiments with synthetic voices using unit selection [1] concatenative synthesis where portions of the database audio recordings are modified for the purpose of producing a wider set of phonemes than is contained in the original voice recordings. Since it is known that performing global signal modification for the purposes of speech synthesis significantly reduces perceived voice quality [2] [3], the modifications that we perform are specifically confined to aperiodic portions of the signal that tend neither to cause concatenation discontinuities nor to convey much of the individual character or affect of the speaker.

We propose three methods to extend the phonetic coverage of unit selection voices (1) by modifying parts of a voice so that extra phones extracted from a donor voice can be added off line; (2) by extending the above methodology by using a harmonic plus noise model (HNM) [4] for speech representation in order to control aspects of the modification; (3) by combining recorded inventories from two voices so that at synthesis time selections can be made from either.

Experiments were conducted to evaluate the strengths and weaknesses of the three methods.

Index Terms: speech synthesis, unit selection, phonetic coverage, unit substitution, Spanish.

1. Introduction

Recently, unit selection concatenative synthesis [1] has become the most popular method of performing speech synthesis. Unit Selection differs from older types of synthesis by generally sounding more natural and spontaneous than formant synthesis or diphone-based concatenative synthesis. Unit selection synthesis typically scores higher than other methods in listener ratings of quality [5]. Building a unit selection synthetic voice typically involves recording many hours of speech by a single speaker. Frequently the speaking style is constrained to be somewhat neutral, so that the synthesized voice can be used for general-purpose applications.

Despite its popularity, unit selection synthesis has a number of limitations. One is that once a voice is recorded, the variations of the voice are limited to the variations within the database. Of course it may be possible to make further recordings of a speaker, but this may not be practical and it may be expensive.

Any techniques that can be used to modify a voice (with the proviso that quality is not degraded) add substantially to the flexibility of unit selection techniques. One such method of extending the range of a voice is to introduce (perhaps limited) prosody modification [2][3]. We would then hope to be able to use the voice for applications where a different prosody, affect, or speaking style is called for.

Voice transformation [6] [7] offers an alternative method of extending the variability of a voice (albeit with the different goal of changing the speaker's individual voice characteristics), but it has not so far produced sufficiently high quality results for use in commercial speech synthesis.

One interesting approach, although taken to reduce database size rather than to expand the range of a voice, is to intermingle natural voice recordings with formant synthesis [8]. The key to this approach is to avoid substitution of highly salient recorded segments by formant synthesis and to only substitute the perceptually less noticeable segments. Replacing selected segments in this way, it was found that the perceived voice quality can remain high, and it was noted that this hybrid synthesis method could allow potentially significant reductions in the size of a database.

The approach we take here is similar, but instead of using formant synthesized segments we use natural segments available from other recorded human voices, and we are interested in *adding* phonemes to a voice database, rather than replacing or substituting them.

2. Applications

We see this work as being potentially useful for applications where a voice may need to be extended in some way, for example to pronounce foreign words. As a specific example, the word "Bush" in Spanish would be strictly pronounced /b/ /u/ /s/ (SAMPA), since there is no /S/ in Spanish. However, in the US, "Bush" is often rendered by Spanish speakers as /b/ /u/ /S/. These loan phonemes typically are produced and understood by Spanish speakers, but are not used except in loan words.

There are languages, such as German and Spanish, where English, French, or Italian loan words are often used. There are also regions where there is a large population living in a linguistically distinct environment and frequently using and adapting foreign names. We would like to be able to synthesize such material accurately without having to resort to adding special recordings. Another problem is that a speaker may be unable to pronounce the required "foreign" phones acceptably, so additional recordings may be impossible.

There are also instances in which the phonetic inventories differ between two dialects or regional accents of a language. In this case, we would like to expand the phonetic coverage of a synthetic voice created to speak one dialect to cover the other dialect as well.

In this paper we implement and evaluate several methods by which such phonetic expansion may be integrated into an already

existing database. Our focus is on Spanish, and specifically on the phenomenon of "seseo," [9] one of the principal differences between European and Latin American Spanish. Seseo refers to the choice between /T/ or /s/ in the pronunciation of words. There is a general rule that in Peninsular (European) Spanish the orthographic symbols z and c (the latter followed by i or e) are pronounced as /T/. In Latin American varieties of Spanish these graphemes are always pronounced as /s/. Thus for the word "gracias"("thanks") the transcription would be /graTias/ in Peninsular Spanish or /grasias/ in Latin American Spanish. Seseo is one major distinction (but certainly not the only distinction) between Old and New World dialects of Spanish.

3. Segment Substitution and Synthesis Methods

We wish to extend the usefulness of a unit selection database by adding units that were not originally present in the voice recordings. Following the observations of [8] we focus in this paper on changing units which carry very little information that could be used to identify an individual speaker. For our experiments, fricatives are among the most interesting of such elements. Specifically, we add /T/ segments to a Latin American Spanish database that contained none, so that the expanded synthetic voice can produce Peninsular Spanish of perceived quality equivalent to the original high quality TTS voice. We use three different methods to achieve that goal.

In all three methods implemented, a general-purpose unit selection database made from a variety of recordings of a female speaker of Latin American Spanish serves as the reference database. The reference database consists of approximately 5 hours of recorded material from a variety of text sources, including news text and interactive prompts.

A second recorded speech database, which we shall refer to as the "donor voice," supplied the loan phonemes by which the reference database was expanded. Both the female speaker and the language (American English) of the donor voice differed from those in the reference database.

3.1. Method 1: Off-line waveform substitution

The first method of modifying the unit selection voice databases that we employ is simple. In this method, waveform segments in the reference database are directly substituted by others from the donor voice, and this segment substitution is performed off-line.

A method was devised to identify segments in the database that could be substituted by a different fricative. Only the /s/ fricatives in the reference database that in Peninsular Spanish would be pronounced as /T/ were substituted. One of the first problems that can arise here is that the unit boundaries in a unit selection database are not always, or even necessarily, on phone boundaries, and so a method is needed that will mark precisely the boundaries of the fricatives of interest, independent of any labeling that exists in the database for the purposes of unit selection synthesis.

In the current experiment, this process was relatively straightforward. The fricatives in question that we chose to examine in detail, /s/ in the reference database and /T/ in the donor voice database, are readily identifiable in a majority of cases by relatively abrupt C-V (unvoiced-voiced) or V-C (voiced-unvoiced) transitions. A method of locating the relevant phone boundaries was derived using a variant of the zero-crossing calculation. Other automatically-marked boundaries were treated with more suspicion, and the entire set of boundaries was manually verified, although with very little modification required. Only a very few segments exhibited possible complications where, for example, the /s/ appeared to be voiced.

In this way, confidence was established in the location of the phone boundaries, both in the reference database and in the set of desired substitute audio material from the donor voice.

Next, the new /T/ audio waveforms from the donor voice were spliced into the reference database in place of the original /s/ audio, with a smooth transition.

With the new audio files and associated phoneme labels, a complete voice was built in the normal fashion and used for unit selection synthesis.

3.2. Method 2: Off-line HNM parameter substitution

A second method is to use a harmonic plus noise model (HNM) [4] representation of speech rather than audio waveforms themselves. In this method the entire database is first converted to HNM parameters. For each frame there is a noise component represented by a set of autoregression coefficients and a set of amplitudes and phases to represent the harmonic component. The HNM parameters were modified, but only the autoregression coefficients were changed, and only when a frame fell time-wise into one of the segments marked for change. In these cases the autoregression coefficients were substituted for a different set derived from the donor voice audio that was substituted directly in method one. The modified set of HNM parameters were then used to synthesize speech. Finally, that speech was used, along with the associated phone labels to build a complete voice suitable for unit selection synthesis.

3.3. Method 3: On-line substitution from combined databases during synthesis

A third method that was explored was to combine the reference and donor voice databases into one. That is, all the database audio files and associated label files for the two different voices were combined. Care was taken to label the phonemes so that there was no overlap of phonetic symbols, except in the case of segments marked as silence, where we felt that a silence in one language sounds much like silence in another. Using these audio files and associated labels a single hybrid voice was built.

Access to the voice can be controlled at the phoneme level, with the choice of phones determining whether we hear one voice in English, or the other voice in Spanish. We were then able to substitute phones simply by specifying a different phone symbol for particular cases, i.e. specifying a /T/ unit rather than a /s/ unit in appropriate instances. Note that in this case there is no attempt made to refine whatever phoneme boundaries were defined in the existing voice database itself. Often these boundary alignments can be less accurate than desired for the purposes of unit substitution.

4. Subjective Evaluation

An experiment was conducted to compare synthesis quality of the above three methods of unit substitution to expand phonetic coverage. The goal was to compare the reference voice (female Latin American Spanish) with four different "hybrid" voices that borrowed /T/ phones from the donor voice (female American English), thus creating synthetic voices that more closely resemble Peninsular Spanish.

4.1. Synthetic voices

Five unit selection synthetic voices, listed below, were used in the experiment.

- **Ref:** The reference female Latin American Spanish unit selection voice.
- **AudHyb:** The hybrid voice described above in Method 1, in which /s/ phones (related to seseo) from the audio database of the reference voice were substituted with audio from /T/ phones taken from the database of the female American English donor voice. All other aspects of the synthesizer, including prosody prediction, were identical to that of Ref, the Latin American reference voice.
- **AEuHyb:** Another Method 1 hybrid voice which differs from AudHyb in that it uses a different prosody module that was developed for European/Peninsular Spanish.
- **HNMHyb:** The hybrid voice described in Method 2, in which HNM parameters rather than audio were substituted.
- **MixHyb:** The hybrid voice described in Method 3, in which the reference and donor voice databases were combined and unit substitution was performed during synthesis.

4.2. Test procedures

A web-based listening test was conducted to measure the subjective quality of each of the five TTS Spanish voices.

Test material consisted of 12 synthetic Spanish sentences randomly selected from a larger set whose durations were all under 6 seconds and with the constraint that each contained at least one instance of a phone affected by seseo. None of the test sentences were represented in the recorded database of the reference voice. Each of the 12 test sentences were synthesized by each of the five TTS voices, yielding a total of 60 test stimuli.

Only adult native speakers of Spanish participated as listeners. The majority of the listeners had no previous experience with synthetic speech, and none were linguists or synthesis specialists. Eight of the ten listeners were native speakers of varieties of Latin American Spanish, while only two were Peninsular Spanish speakers. The unequal representation of the two varieties of Spanish is a flaw of the experiment that we hope to correct given more time to locate Peninsular Spanish speakers.

Listeners were asked (all instructions were printed on the website in Spanish) to click an icon to listen to a test file. They could listen as many times as they wished. They then rated the speech quality of the file along a five-point scale: (1) Pésimo (Bad), (2) Malo (Poor), (3) Regular (Fair), (4) Bueno (Good), (5) Excelente (Excellent). The order of test stimuli was randomized independently for each listener. Before beginning the test, five practice stimuli (one for each TTS voice tested) were presented and rated in order to familiarize listeners with the procedure and the range of stimuli they would hear, and also to allow them to adjust their preferred audio level in advance of the test. All files were equivalent in level. The tests were conducted in relatively quiet individual office settings. Three listeners reported using headphones, and 7 used speakers. The test typically took from 15 to 20 minutes to complete.

4.3. Results

Mean ratings for each of the TTS voices are listed in Table 1.

TTS Voice	M.O.S.	Std.Error
Ref	3.775	.155
AudHyb	3.642	.136
AEuHyb	3.633	.129
HNMHyb	3.575	.116
MixHyb	3.367	.112

Table 1: *Mean opinion scores of the TTS Voices.*

A repeated measures Analysis of Variance (ANOVA) was performed on the rating data collected (600 observations). The ANOVA design was TTS(5) + Sentence(12) + TTS * Sentence (60). Once more Peninsular Spanish listeners have participated in the test, the ANOVA design will also include a Group(2) (between-listener) factor of Spanish dialect group.

There was a main effect of TTS ($F(4,36)=3.58$, $p<0.015$), indicating significant differences in ratings between TTS voices. Pairwise comparisons indicated that there was no significant difference in ratings between the three highest rated TTS voices, Ref, AudHyb, and AEuHyb, but Ref ratings were significantly higher than HNMHyb and MixHyb voices.

Results of the ANOVA also showed a main effect for Sentence ($F(11,99)=6.417$, $p<0.0001$), indicating that across all TTS voices, ratings among sentences differed significantly. There was also a significant TTS * Sentence interaction ($F(44,396)=4.130$, $p<0.0001$), because ratings for individual sentences differed among TTS voices.

Although because of the small and unbalanced number per dialect of Spanish listeners, no statistics could be performed to test the effect of native dialect, even the differences observed so far are suggestive that native dialect influences subjective quality ratings. The highest rated TTS voice for Latin American listeners was Ref (MOS = 3.823), the only "pure" Latin American TTS voice tested. On the other hand, for European Spanish listeners, AudHyb (MOS = 3.792) was the most highly rated TTS voice, while Ref scored over 0.2 lower (MOS = 3.583).

5. Discussion

On the basis of the high ratings achieved by AudHyb and AEuHyb TTS voices in the subjective evaluation, TTS quality does not appear to be affected adversely by unit substitution of carefully verified and selected audio from another voice and language.

The Peninsular Spanish module used for prosody prediction in AEuHyb did not appear to affect overall ratings of subjective quality for the unit selection TTS voice tested. A similar result was observed with prosodically unmodified unit selection synthesis in English [3].

The hybrid voice that substituted HNM parameters rather than audio was slightly less successful, but since there was no reference condition that used HNM representation without substitution, it is unclear whether the slightly lower mean opinion score was related to unit substitution or simply the parameterization itself.

The relatively poor rating of the MixHyb voice reveals the importance for unit substitution of the careful verification of phone boundaries that was performed for the other three hybrid TTS voices. MixHyb's use, for the purposes of unit substitution, of the same automatically labeled and aligned phone boundaries that are used for standard synthesis resulted in poorer quality synthesis

than AudHyb and AEHyb.

6. Conclusions

At least one of the unit substitution methods presented in this paper represents a viable method of modifying a synthetic voice in a way that adds flexibility and does not noticeably damage the quality of the resulting signal. We think that the other two methods, though they currently produce slightly lower quality synthesis, remain promising techniques nevertheless.

We intend to extend these methods and use them in our synthesizer. We also intend to look at more challenging cases involving voiced consonants and are interested in studying what (including prosody [10]) is involved in changing from one dialect to another.

7. References

[1] Hunt, A. and Black, A. "Unit selection in a concatenative speech synthesis system using large speech database", ICASSP, 373-376, 1996.

[2] Beutnagel, M., Conkie, A.,and Syrdal, A. K. "Diphone synthesis using unit selection", Third ESCA Speech Synthesis Workshop, Jenolan Caves, Australia, Nov. 1998, 185-190.

[3] Jilka, M., Syrdal, A. K., Conkie, A., and Kapilow, D. "Effects on TTS quality of realizing natural prosodic variations", ICPhS, Aug. 2003, 2549-2552.

[4] Stylianou, Y., Laroche, J., and Moulines, E. High-Quality Speech Modification based on a Harmonic + Noise Model", Eurospeech, Madrid, Spain, 1995, 451-454.

[5] Vazquez Alvarez, Y. and Huckvale, M. "The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems", ICSLP, Denver, Sept. 2002, 329-332.

[6] Lee, K.-S., Youn, D. H., and Cha, I. W. "A new voice transformation method based on both linear and nonlinear prediction analysis",ICSLP, 1996, Vol. 3: 1401-1404.

[7] Stylianou, Y., Cappé, O., and Moulines, E. "Continuous Probabilistic Transform for Voice Conversion", IEEE Trans. Speech and Audio Proc., 6(2):131-142, 1998.

[8] Hertz, Susan R., "Integration of rule-based formant synthesis and waveform concatenation: a hybrid approach to text-to-speech synthesis", IEEE 2002 Workshop on Speech Synthesis, Santa Monica, CA, Sept. 2002.

[9] Navarro Tomás, T., Manual de Pronunciación Española, 20th edition. Madrid: CSIC, 1980.

[10] Jilka, M. "The Contribution of Intonation to the Perception of Foreign Accent", Doctoral Dissertation, Arbeiten des Instituts für Maschinelle Sprachverarbeitung (AIMS) Vol. 6(3), University of Stuttgart, 2000.