

Unit-Selection Text-to-Speech Synthesis Using an Asynchronous Interpolation Model

Alexander Kain^{1,2}

Jan P. H. van Santen^{1,2}

¹ Center for Spoken Language Understanding (CSLU)
OGI School of Science & Engineering at OHSU
20000 NW Walker Road, Beaverton, OR 97006, USA

² BioSpeech, Inc.
940 Upper Devon Lane
Lake Oswego, OR 97034, USA

Abstract

We describe the Asynchronous Interpolation Model, which represents speech as a composition of several different types of feature streams that are computed using asynchronous interpolation of neighboring basis vectors, according to transition weights. When applied to the acoustic inventory of a concatenative Text-to-Speech synthesizer, the model eliminates concatenation errors and affords opportunities for high rates of compression and voice transformation. We propose a particular instance of the model that uses formant frequency values and formant-normalized complex spectra as two types of streams, in conjunction with a unit-selection synthesizer. During analysis, basis vectors and transition weights were estimated automatically, using three different labeling schemes and dynamic programming methods. An evaluation of the intelligibility and quality of the synthesized speech showed significant improvements over a standard, size-matched compression scheme. The proposed method was also able to convincingly transform speaker characteristics through replacement of basis vectors.

1. Introduction

Today's most natural sounding Text-to-Speech (TTS) synthesis systems are based on the *concatenative synthesis* approach, which uses a multitude of pre-recorded speech "chunks" (a contiguous section of natural speech) of a single speaker, stored in an *acoustic inventory*, to stitch together a new output signal. The quality of the resulting speech relates directly to the size of the database, because the larger the chunks, the fewer the number of concatenation points at which audible artifacts can occur. Moreover, when the prosodic space is not covered by the acoustic inventory, prosodic modification becomes necessary, further degrading the speech signal. The concatenative approach can be contrasted with the *formant synthesis* approach, which is compact in size, gives full prosodic and spectral control over the speech signal, and is highly intelligible, but which does not sound very natural.

Researchers have attempted to improve the problem of audible discontinuities in concatenative synthesis, by interpolating in the formant, waveform, or suitable linear predictive coding domains [1, 2, 3, 4]. However, these approaches commonly neither increase synthesis flexibility nor address the issue of compactness.

We propose a model that combines aspects of both the formant and the concatenative approaches, called the Asynchronous Interpolation Model (AIM). Its features are:

- Elimination of concatenation errors, because speech

units of the acoustic inventory have identical representations at concatenation points.

- Opportunity for compression. Even though memory is continuing to decline in price and increase in capacity, it is attractive to control the size/quality trade-off, and thus enable large acoustic inventories on (extremely) storage-limited devices such as cellphones. AIM can take advantage of the special properties of an acoustic inventory, which are that the inventory consists of a single speaker, is acoustically constant and noise-free, non-real-time encoding is possible, all data is known beforehand, and additional information such as phonetic content is available.
- Increased spectral flexibility. For example, changing the duration of a segment of speech changes its spectral properties in complex ways. The increased flexibility of AIM allows non-linear, independent changes of different aspects of the speech signal.
- Voice transformation with a small number of required samples from the target speaker, making it possible to easily produce additional voices from an existing acoustic inventory, as opposed to recording an entire new inventory for the voice, which is time-consuming, tedious, and expensive. Example applications include systems for persons with voice disorders who use TTS synthesizers to communicate. Many such people can, with great effort, produce clear speech intermittently, which can then be used as training samples, ultimately rendering the output of their TTS system with their own voice.

In previous work, we have applied AIM to a diphone synthesizer, reducing the size of a 6.5 MB inventory to 57 kB (1:114 compression) at 8 dB spectral distortion, while eliminating concatenation errors [5]. In this paper, we extend our work to a unit-selection TTS synthesizer, which leads to new approaches during analysis and synthesis. Section 2 introduces the core ideas, as well as the general and implementation-specific forms of AIM. Sections 3 and 4 describe the analysis and synthesis of speech under the model. Section 5 evaluates the TTS system with respect to its intelligibility, quality, and speaker recognizability. Section 6 concludes the paper and discusses future directions.

2. The Asynchronous Interpolation Model

The core idea of AIM is to represent a short region (on the order of 5–10 ms) of speech as a *composition* of several types of features called *streams*. Each stream is computed by asynchronous

interpolation of neighboring *basis vector* features. Each basis vector is associated (labeled) with a particular phoneme, allophone, or more specialized unit and may contain additional information about phonetic and prosodic context. Thus, the speech region is described by the varying degrees of influence of several types of preceding and following acoustic features. In this section, we extend and improve upon the notation reported previously [6].

Representing speech as an interpolation between vectors has been researched before; for example, the temporal decomposition approach [7, 8] decomposes speech into arbitrary event targets that describe successive events. Our method stands apart in that the phonemic identities of the basis vectors are known, and asynchronous interpolations are carried out on several streams consisting of different types of features.

2.1. General Form

Given a speech waveform, let the complex spectrum \mathbf{X} at frame m be equal to a composition operation \mathcal{C} on the values of N streams \mathbf{s} at that frame

$$\mathbf{X}[m] = \mathcal{C}(\mathbf{s}_1[m], \dots, \mathbf{s}_N[m]) \quad (1)$$

where different streams represent different types of feature trajectories. An individual stream is calculated by the interpolation

$$\mathbf{s}_n[m] = \sum_{k=1}^K w_n^{U_k}[m] \cdot \mathbf{b}_n^{u_k} \quad (2)$$

where $\mathbf{b}_n^{u_k}$ are the *basis vectors* associated with stream n and acoustic event u_k , and $w_n^{U_k}[m]$ are the *transition weights* at frame m that are associated with stream n and context

$$U_k = u_{k-l}, u_{k-l+1}, \dots, u_{k-1}, u_k, u_{k+1}, \dots, u_{k+r-1}, u_{k+r}$$

that includes the l previous and r following acoustic events. The summation is performed over K acoustic events. In addition, for a given frame m and stream n , transition weights are constrained by

$$\sum_{k=1}^K w_n^{U_k}[m] = 1 \quad (3)$$

to ensure a convex operation. When choosing speech features, care must be taken that they are “interpolatable” so that stream values are valid in a physical sense at all times; for example, formant parameters are interpolatable, but polynomial filter coefficients are not.

2.2. Implementation

In our specific implementation, we reduced phonetic and prosodic context by constraining the summation of Equation 2 to only depend on the previous and the next unit; in other words, the influence of a basis vector never extends beyond its neighbor. We chose two types of features, namely formant frequency locations and the formant-normalized complex spectrum. The latter is the result of modifying the complex spectrum so that formants appear at constant neutral values, allowing the interpolation of spectra without adding extraneous formants. Therefore, Equation 1 becomes

$$\mathbf{X}[m] = \mathcal{C}(\mathbf{s}_s[m], \mathbf{s}_f[m]) \quad (4)$$

where the subscripts refer to the association with spectral and formant information, respectively. The composition operator

\mathcal{C} was implemented as a non-linear warping of the formant-normalized spectral feature stream to obtain a spectrum with formants at the locations specified by the formant stream (more on this in Section 2.2.2).

The reduced context allows combining Equations 2 and 3, resulting in

$$\begin{aligned} \mathbf{s}_s[m] &= w_s^{u_l \rightarrow u_r}[m] \cdot \mathbf{b}_s^{u_l} + (1 - w_s^{u_l \rightarrow u_r}[m]) \cdot \mathbf{b}_s^{u_r} \\ \mathbf{s}_f[m] &= w_f^{u_l \rightarrow u_r}[m] \cdot \mathbf{b}_f^{u_l} + (1 - w_f^{u_l \rightarrow u_r}[m]) \cdot \mathbf{b}_f^{u_r} \end{aligned} \quad (5)$$

where u_l and u_r are acoustic events left and right of frame m , and m varies from the frame associated with event u_l to the frame associated with u_r .

Our choice of features was guided by the observation that in transitions between most phonemes, formant frequencies and the overall spectral shape change asynchronously (although this instance of the model makes the simplifying assumption that the formants themselves are synchronous). For example, a transition from /i:/ to /v/, as in the word “leave”, shows a change in formants that starts well before the onset of frication. Another view is to regard the resulting system as an equivalent to image morphing, where salient features are used to mark important regions of two still images, and transitions are created by smoothly moving the salient features while modifying the underlying still images appropriately. In our case we used formants as salient features to render a good approximation of the transition between two sounds, which could not be achieved by a simple cross-fade.

2.2.1. Basis Vector Labeling

We selected basis vector label names similar to the Worldbet [9] phonetic labels for American English. Since basis vectors represent single acoustic events, some phonemes needed to contain several basis vectors. Specifically, diphthongs contained two separate basis vectors for the two different targets (/aI/: “aI1”, “aI2”), voiced plosives contained two basis vectors for closure and burst (/b/: “bc”, “b”), and unvoiced plosives contained three basis vectors for closure, burst, and aspiration (/t/: “tc”, “tb”, “th”). Finally, we represent affricates as a combination of other basis vectors (/tS/: “tc”, “tb”, “S”).

Two different basis vector occurrences with the same label in the acoustic inventory can be treated as identical or distinct. This gave rise to the following three labeling schemes:

Global In the global labeling scheme all basis vectors with the same label were shared, resulting in typically less than 60 basis vectors. This led to the smallest representation of the acoustic inventory and thus also gave the highest compression rate.

Local The opposite of the global scheme, the local scheme considers every basis vector in the inventory as unique. Special care must be taken during synthesis when concatenating two units with distinct basis vectors at the cut-point to ensure smoothness (see Section 4). This scheme still provided a high rate of compression because the majority of frames are within transitions and are represented by transition weights only.

Automatic This scheme allowed the selection of an arbitrary size or quality criterion (as specified by an objective function) on the continuum spanned by the two previous schemes. This was implemented either by growing the global scheme and iteratively splitting and reassigning shared basis vectors, or by pruning the local scheme and iteratively merging two unique basis vectors.

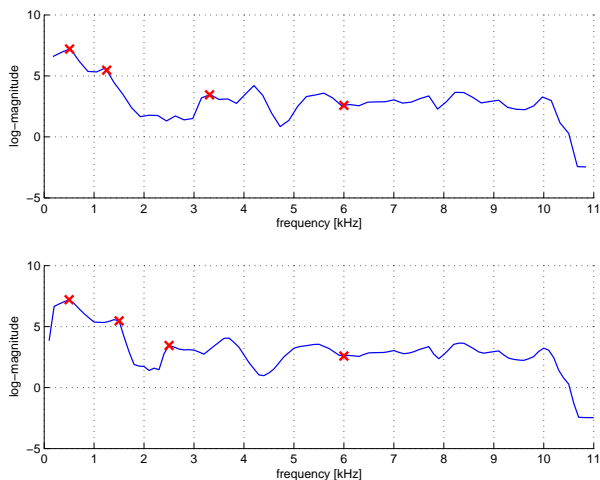


Figure 1: The effect of the composition operation. Given a log-magnitude spectrum of the phoneme /l/ with original frequency locations (top), the composition operation creates a non-uniformly resampled version to align with the desired frequency locations (bottom). Formant frequencies F1, F2, and F3, as well as the modification-cutoff frequency are located at the markers.

2.2.2. Composition Operation

The task of the composition operation is to receive a vector of stream values and to then render a short segment of speech. In our case the inputs are formant-normalized complex spectra and formant frequency values, and the composition consists of returning a modified complex spectrum with the neutral formant frequency locations changed to the specified ones.

Modifying formant frequencies in the natural spectrum has been previously researched [10, 11, 12]. Our implementation consists of non-uniformly resampling the original spectrum (see Figure 1). In addition to formant frequencies, we specify a modification-cutoff frequency at 6000 Hz to stop modification of the spectrum at and above that frequency. Conversely, the formant-normalized spectra themselves were initially created by modifying the original spectrum with associated original formant frequency locations to have formants at a constant neutral location.

3. Analysis

During analysis, synthesis, and evaluation, the system utilizes a small unit-selection database of a female speaker “AS” [13], which covers all diphones and specific triphones that are known to have a significant amount of coarticulation, but which does not have complete prosodic coverage.

3.1. Basis Vectors

In the proposed implementation, basis vectors contain information about both the complex spectrum and formant frequency locations. Therefore, the analysis process begins by making initial estimates of formant frequency trajectories F1, F2, and F3, using the ESPS `get_formant` algorithm [14].

The locations of basis vectors relative to phoneme boundaries are initialized as follows: When the phoneme contains just one basis vector, its location is set to that point which will, on average, result in the smallest concatenation error. For

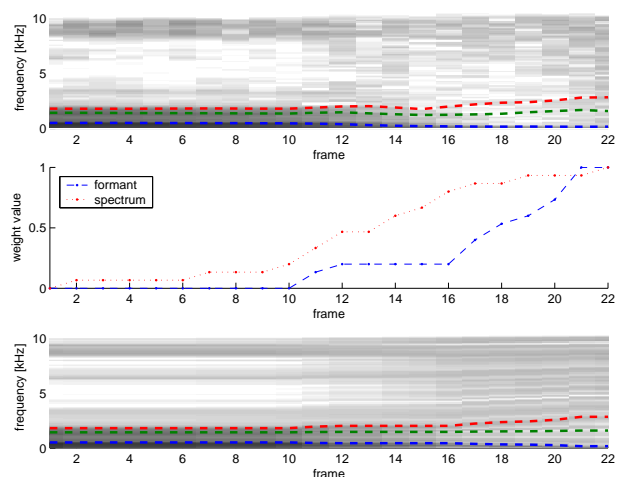


Figure 2: Transition weight analysis. The top panel shows the original log-magnitude spectrogram for the transition between /9r/ and /v/, with the original formant frequency trajectories superimposed. The middle panel shows the resulting weights after analysis. The asynchronous nature of the weights is easily observable. The bottom panel shows a resynthesis of the transition using the previously analyzed basis vectors and transition weights.

phonemes with two or more targets simple heuristics are employed, such as assigning the second basis vector at the 80% point of the total duration of a diphthong.

Both basis vector locations and formant frequency trajectories were manually corrected using a standard labeling tool in conjunction with a pen input device. This proved especially necessary in the following two cases: (1) to fine-tune the location of basis vectors during stops, affricates, and diphthongs, and (2) to create appropriate formant frequency locations in regions in which formants were not clearly visible, such as during a closure preceding a stop. In the latter case, formant frequencies were assigned in accordance with locus theory [15].

To extract complex spectra, we perform a pitch-synchronous sinusoidal analysis over two frames nearest to the basis vector location and store the magnitude and phase of each harmonic sinusoid, as well as fundamental frequency and voicing information of the analysis frame.

3.2. Transition Weights

For each transition, we fit the transition weights by first assuming a straight-line transition $w = 0, 1/Q, \dots, (Q-1)/Q$, where Q represents the weight value resolution; for example, we use $Q = 16 = 2^4$ which allows weights to be stored in 4 bits. Then, the formant-normalized magnitude spectral stream and formant frequency stream are constructed using local basis vectors and the straight-line weights. (The phase spectrum is ignored during fitting.) Finally, the streams are separately aligned to the original formant and spectral transitions using a dynamic time warping (DTW) algorithm (see Figure 2). In cases where original formant trajectories are unavailable, a joint DTW can be used [6].

The DTW algorithm has local constraints that insure monotonically increasing transition weights. There are no global constraints and the local constraints allow for maximally discontinuous changes in the weights from one frame to the next. This is

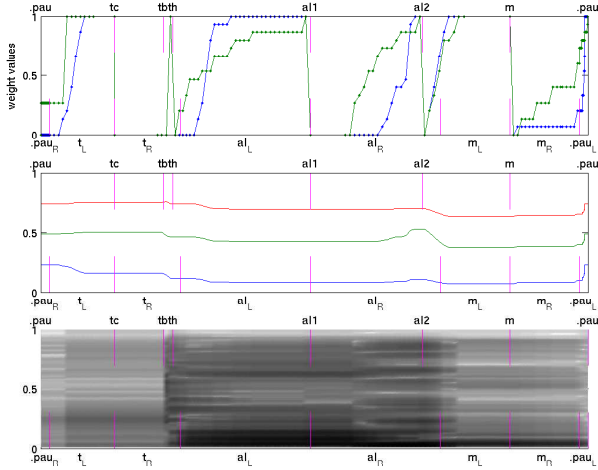


Figure 3: Basis vectors and transition weights used for synthetic utterance (top panel). Basis vectors “tc”, “al1”, and “m” are surrounded by weight values of ones (on their left) or zeros (on their right), respectively, due to synthetic lengthening of the units. The formant and spectral streams are displayed as trajectories (middle panel) and the formant-normalized log-magnitude spectrogram (bottom panel). Lines at the tops of panels mark the position and identity of a basis vector, whereas lines at the bottoms of panels denote the diphone boundaries, their identities labeled at the center.

needed because many transitions are quite abrupt (for example, nasal to vowel transitions).

Transition weight trajectories could be further regularized by replacing them with parametric functions, for example a sigmoidal function. Moreover, weight trajectories of certain classes of similar transitions (for example vowel to nasal transitions) could be tied to a single model. Both of these optional steps would yield additional storage savings.

4. Synthesis

The compressed acoustic inventory consists of two 4-bit weights, $w_s^{u_l \rightarrow u_r}[m]$ and $w_f^{u_l \rightarrow u_r}[m]$, for each speech frame, and an associated basis vector time and identity list that references the collection of basis vectors \mathbf{b}_s^u and \mathbf{b}_f^u , in addition to the traditional list of units that are used during the acoustic inventory search. During synthesis, we first construct basis vectors and transition weights for the synthetic utterance, by assigning weights, basis vector locations and basis vector identities according to the output of the unit search and the specified synthetic durations. When synthetic durations are shorter than the original units, the unit is shortened by compressing the times at which weights and basis vectors occur. When synthetic durations are longer than the original units, we leave the original weight trajectories unmodified, but instead shift the weights left when we are synthesizing the left side of a phoneme, and shift the weights right for the right side of a phoneme. The resulting effect is that phonemes are lengthened at their centers during stretching, but that the transitions themselves are at their original speeds (see Figure 3).

When two non-identical basis vectors fall onto the same point in time, we merge the basis vectors by taking their average; thus creating a new, temporary basis vector (by definition, streams are interpolatable). This situation occurs when the local

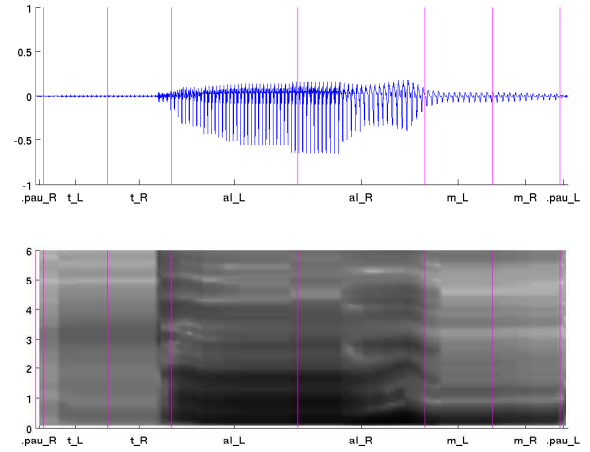


Figure 4: Synthetic waveform (top panel) and pitch-synchronous log-magnitude spectrogram (bottom panel) with diphone boundary lines.

or automatic labeling scheme is used and we are concatenating across two basis vectors associated with the same phoneme, but from two different contexts.

After constructing basis vectors and transition weights, Equations 5 and then 4 are used to calculate the complex spectrogram of the synthetic utterance, which is finally rendered as a waveform by a pitch-synchronous sinusoidal synthesis algorithm (see Figure 4).

AIM also allows a new approach to the spectral aspect of voice transformation, by regarding basis vectors as speaker-dependent, but transition weights as speaker-independent. Using the global labeling scheme described in Section 2.2.1, we estimated a small number of basis vectors for several new target speakers. Then, transformed speech is produced by using the original speaker’s transition weights with the desired target speaker’s basis vectors.

5. Evaluation

5.1. Intelligibility and Quality

The following four conditions were compared: (1) the standard OGI TTS baseline system [13] at 352.8 kbps, (2) the baseline compressed with the Speex CELP coder [16] at 8.0 kbps, (3) the baseline compressed with the Speex CELP coder at 3.4 kbps, and (4) the BioSpeech AIM TTS system using the global labeling scheme at 3.4 kbps. The average bit rate for AIM was computed as follows: Given 54 basis vectors with an average dimension of 62, where each component is represented by 16 bits, yields 53,568 bits. Each of the 63,716 frames of the acoustic inventory contains an 8-bit number that marks the position of the frame; in addition, each frame contains two 4-bit transition weights, for a total of 1,019,456 bits. Finally, the 132,300-bit wave library is added, for a grand total of representing the database in 1,205,324 bits or 3,414 bps. Compared to the original representation of 124,530,928 bits, or 352.8 kbps, this represents a 103:1 compression rate.

The text material used in these experiments consisted of 48 sentences, randomly selected from the IEEE Harvard Psychoacoustic Sentences [17], containing five keywords each (e. g. “His shirt was clean but one button was gone”). Each sentence

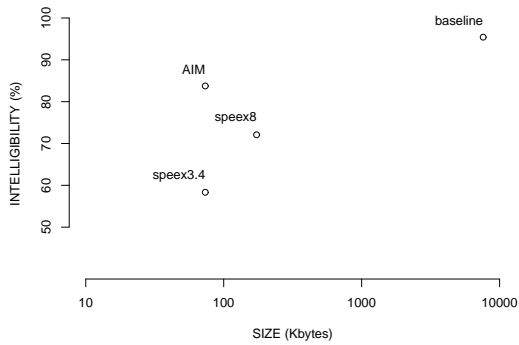


Figure 5: Word intelligibility, defined as the percentage of key-words correctly repeated per sentence.

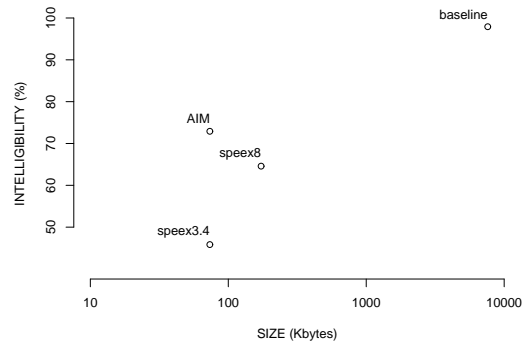


Figure 6: Sentence intelligibility, defined as the percentage of sentences correctly repeated in their entirety.

was synthesized in each of the four conditions.

Six listeners aged 24–35 participated, all native speakers of English and unfamiliar with the goals of the study. Listeners heard an utterance exactly once, attempted to repeat the utterance, and then rated its speech quality on a 1–5 Mean Opinion Score (MOS) scale (“bad”, “poor”, “fair”, “good”, “excellent”). A test administrator scored the number of key words that were repeated correctly, while the rating was recorded automatically. The test was designed so that condition and presentation order were uncorrelated; therefore any effects due to condition cannot be attributed to some conditions being presented relatively late (or early) in the experiment.

Figures 5 and 6 show the results for word intelligibility (I_W), defined as the percentage of keywords correctly repeated per sentence, and sentence Intelligibility (I_S), defined as the percentage of sentences correctly repeated in their entirety. Figure 7 shows quality (Q) represented by the mean opinion score, averaged over all listeners and all sentences in that particular condition. Statistical tests (planned t -tests) indicated that AIM was significantly superior in intelligibility and quality to the size-matched 3.4 kbps coder condition (I_W : $p < 0.005$; I_S : $p < 0.015$; Q : $p < 0.001$). AIM was also superior in both ways to the larger 8 kbps coder condition, but this was significant only for quality (Q : $p < 0.001$).

5.2. Speaker Recognizability

In this test, a source speaker’s basis vectors of an acoustic inventory were replaced with basis vectors from a target speaker’s acoustic inventory, while leaving the transition weights unchanged. Prosody was kept exactly constant for all stimuli to ensure that speaker recognizability performance was measured based on spectral cues only, and not on prosodic cues.

The text material used in this experiment consisted of 40 sentences, randomly selected from the IEEE Harvard Psychoacoustic Sentences [17]. The sentences were synthesized using AIM with representations derived from the acoustic inventory of five male voices, aged 21–39, and whose native language was American English. The local labeling scheme was used for highest synthesis quality. For 20 of the sentences, the original basis vectors were replaced by basis vectors derived from exactly one of the other four voices.

A speaker recognizability test was chosen to evaluate voice transformation performance [18]. During testing, six listeners

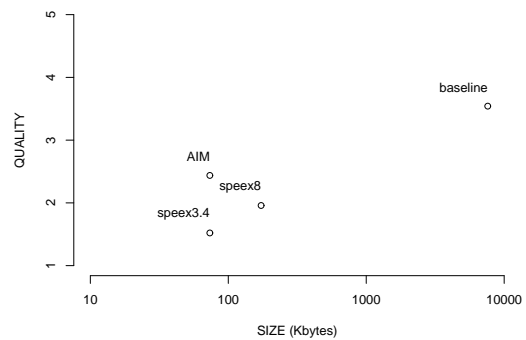


Figure 7: Mean Opinion Score (MOS) of speech quality, on a 1–5 scale (“bad”, “poor”, “fair”, “good”, “excellent”).

(the same as in Section 5.1) heard two utterances in sequence. One of them was the voice transformation condition described above, and the other was the normally synthesized version of a different sentence of either the same speaker or a different speaker. The task was to decide whether the two utterances were from the same speaker or from two different speakers. The response alternatives were “definitely different”, “kind of different”, “unsure”, “kind of same”, “definitely same”, and were recorded automatically. Note the equal number (20) of correct “same” and “different” responses.

All six listeners had higher percentages of matching speakers when they indicated “same” compared to “different”, significant at $p < 0.025$ using a 1-tailed Sign test. Except for one listener, all speakers showed a completely monotonically decreasing response pattern, as shown in Figure 8. Four out of six listeners recognized speakers as being the same 100% correctly when they were certain of the speakers being the same. Conversely, three out of four speakers recognized speakers as different 100% correctly when they were certain of their choice. This indicates that, even when no prosodic cues are available and different sentences are presented, the AIM method preserves adequate speaker information to enable listeners to determine speaker identity.

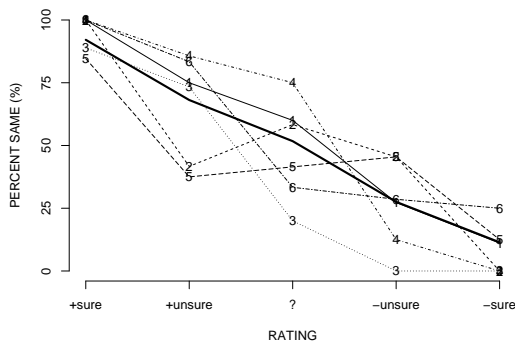


Figure 8: Speaker Recognizability, represented by percentages of items where the two speakers of the stimuli presented are the same, as a function of listener rating. "+" and "-" indicate same and different, respectively. For example, "-sure" refers to "definitely different". Listener numbers are shown on the curves; the heavy curve represents the mean over all listeners.

6. Conclusion

We have described a speech synthesis system based on the Asynchronous Interpolation Model, which represents speech as a composition of several streams that are computed using asynchronous interpolation of neighboring basis vectors. Applied to a concatenative TTS system's acoustic inventory, the model avoids concatenation errors during synthesis, and affords opportunities for variable compression and a new approach to voice transformation. During evaluation, AIM produced significantly higher quality and intelligibility than speech that has been compressed by traditional methods, using sizes equal to AIM or more than twice as large as AIM. The AIM compression ratio in this study was 103:1; this could easily be further increased by further parametrization of transition weights. Results also showed that AIM produces speech that can be reliably identified with a desired target speaker, using an extremely small set of training speech.

Further enhancements are necessary to increase intelligibility and quality scores. One of these would be a more sophisticated method of formant manipulation, which currently was implemented using a simple frequency warping. Another enhancement would be to model the deterministic and stochastic part of a speech frame separately, allowing for higher quality modeling of noise when a single spectral basis vector is repeated several times throughout a transition. Finally, we plan on investigating approaches that will automatically insert additional basis vectors, thus enabling a complete reconstruction of the original acoustic inventory in the limit.

7. Acknowledgments

Part of this research was funded by United States National Science Foundation STTR grant 0441125. Oregon Health & Science University, Dr. Kain, and Dr. van Santen have a significant financial interest in BioSpeech, Inc., a company that may have a commercial interest in the results of this research and technology. This potential conflict was reviewed and a management plan approved by the Conflict of Interest in Research Committee and the Integrity Program Oversight Council was implemented.

8. References

- [1] H. Mizuno, M. Abe, and T. Hirowaka, "Waveform-based speech synthesis approach with a formant frequency modification," in *ICASSP*, 1993, pp. 195–198.
- [2] J. Wouters and M. Macon, "Control of spectral dynamics in concatenative speech synthesis," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 1, pp. 30–38, Jan. 2001.
- [3] D. T. Chappell and J. H. L. Hansen, "A comparison of spectral smoothing methods for segment concatenation based speech synthesis," *Speech Communication*, vol. 36, no. 3, pp. 343–373, 2002.
- [4] P. H. Low, C. H. Ho, and S. Yaseghi, "Using estimated formant tracks for formant smoothing in text to speech synthesis," in *ASRU*, 2003, pp. 688–693.
- [5] A. Kain and J. van Santen, "Compression of acoustic inventories using asynchronous interpolation," in *IEEE Workshop on Speech Synthesis*, 2002, pp. 83–86.
- [6] A. Kain and J. van Santen, "A speech model of acoustic inventories based on asynchronous interpolation," in *EUROSPEECH*, 2003, pp. 329–332.
- [7] B. Atal, "Efficient coding for LPC parameters by temporal decomposition," in *ICASSP*, 1983, pp. 81–84.
- [8] S. Ghaemmaghami, M. Deriche, and B. Boashash, "Comparative study of different parameters for temporal decomposition based speech coding," in *ICASSP*, 1997.
- [9] J. Hieronymus, "ASCII phonetic symbols for the world's languages: Worldbet," Tech. Rep., Bell Labs, 1993.
- [10] Y.-S. Hsiao and D.G. Childers, "A new approach to formant estimation and modification based on pole interaction," in *Thirtieth asilomar conference on signals, systems and computers*, 1996, vol. 1, pp. 783–787.
- [11] R. W. Morris and M. A. Clements, "Modification of formants in the line spectrum domain," *IEEE Signal Processing Letters*, vol. 9, pp. 19–21, Jan. 2002.
- [12] E. Turajlic, D. Rentzos, S. Vaseghi, and C.-H. Ho, "Evaluation of methods for parametric formant transformation in voice conversion," in *ICASSP*, 2003, pp. 724–727.
- [13] M. Macon, A. Cronk, J. Wouters, and A. Kain, "OGIresLPC: Diphone synthesizer using residual-excited linear prediction," Tech. Rep. CSE-97-007, Dept. of Computer Science, Oregon Graduate Institute of Science and Technology, Portland, OR, Sept. 1997.
- [14] Entropic Research Laboratory, "Entropic Signal Processing System (ESPS) Waves+," Software, Aug. 1993.
- [15] D. Broad and F. Clermont, "A methodology for modeling vowel formant contours in CVC context," *JASA*, vol. 81, no. 1, pp. 155–165, Jan. 1987.
- [16] J.-M. Valin, "Speex: A Free Codec For Free Speech," www.speex.org, 2006.
- [17] E. H. Rothaus, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silberger, G. E. Urbaneck, and M. Weinstock, "IEEE Recommended practice for speech quality measurements," *IEEE Trans. on Audio Electroacoustics*, vol. 17, pp. 227–246, 1969.
- [18] A. Kain, *High Resolution Voice Transformation*, Ph.D. thesis, OGI School of Science & Engineering at Oregon Health & Science University, 2001.