Incorporating Duration and Intonation Models in Filipino Speech Synthesis

Lito Rodel S. Lazaro, Leslie L. Policarpio, and Rowena Cristina L. Guevara
Digital Signal Processing Laboratory, Electrical and Electronics Engineering Institute
University of the Philippines Diliman
lslazaro1@up.edu.ph, llpolicarpio@up.edu.ph, gev@eee.upd.edu.ph

***Abstract*—In this paper we describe the development of an intonation model and a duration model to generate prosody for the Filipino language. Z-scores of normalized durations are used for the duration model and the Tilt parameters are used for the intonation model. The Filipino Speech Corpus (FSC) is the source of statistical data for modeling the duration and intonation. A Classification and Regression Tree (CART) generator is used to build the model for duration and intonation.**

The Harmonic plus Noise Model (HNM) is developed for the FSC. The diphones are concatenated to produce the synthetic speech and HNM is used to modify the prosody.

The synthesized speech is evaluated using the Mean Opinion Score (MOS). Results show that the duration model and the intonation model needs improvement. HNM synthesis performs slightly better than TD-PSOLA (time-domain pitch synchronous overlap-add).

***Index Terms*—Filipino, Speech Synthesis, Prosody, HNM**

I. Introduction

Speech synthesis is the automatic generation of speech waveforms with the use of a machine. Speech is the primary means of communication between people; hence, speech synthesis is a natural way for man-machine interface.

Synthetic speech may be used in several applications. An important application of speech synthesis is in reading and communication aid for the handicapped. With speech synthesis, digital talking books which will only require text inputs are now available. Synthesized speech gives the deaf and vocally handicapped an opportunity to communicate with people who do not understand sign language. Synthesized speech can also be used as an educational tool for special tasks such as spelling and teaching pronunciation for different languages. It can also be used with interactive educational applications. Synthetic speech can also be used in interactive games, automated call centers and hand-held devices for quick information.

Initially, the research in speech synthesis is focused in producing intelligible speech. Current synthesizers are now able to produce intelligible speech, and the new goal is the improvement of naturalness of synthesized speech. A good model of prosody is necessary for synthesized speech to sound natural.

Several Text-to-Speech synthesis systems of the Filipino language have been developed in the UP Digital Signal Processing (DSP) Laboratory. All implementations used the time domain pitch synchronous overlap-add (TD-PSOLA) method.

Corpus and Liampo [1], in 2001, developed a continuous text-to-speech synthesizer of Tagalog words with 20 phonemes of the language used as basic acoustical units. In this project, only the intonation part of the prosody generation was put to use. They used punctuation marks at the end of sentences to determine the appropriate intonation pattern. The prosody evaluation test for the Tagalog sentences produced unsatisfactory results.

In 2002, prosody development for Filipino Text-to-Speech systems [2] was done. In line with this, the Filipino Speech Corpus (FSC) was built. The FSC was used as the source of data for the project. Isolated words were used as concatenation units, and paragraphs and sentences in the corpus were used as source of prosody. Dynamic Time Warping (DTW) was used to align the pitch contour and produce the duration contour. Acceptability test shows that the produced synthesized speech is comparable to commercially available Text-to-Speech (TTS) systems. The system was used in in 2007 to generate the synthetic speech of the English text translated to Filipino [3].

A concatenative synthesis of two-syllable Filipino words [4] was developed to address the problems associated with the Tagalog TTS Synthesizer. Filipino words were characterized to differentiate words of the same spelling but different meaning, depending on the pronunciation. The characterization took note of the pitch and duration of utterance of each word. The synthesizer accepts two-syllable text and outputs all the possible utterances for that word. Discontinuities at concatenation points degrade the quality of the synthesized word.

Much of the prosody generation used in the previous Filipino TTS systems rely on a corpus of phrases which prohibits the synthesis of more phrases. In April 2005, an automatic means of modeling the duration of Filipino [5] was developed. Syllables were used as acoustic units. The duration of the generated synthetic speech was acceptable but the syllable concatenation using TD-PSOLA needs to be refined for commercial acceptability.

To be able to produce a natural-sounding synthetic speech, prosody must be incorporated. Automation of the prosody generation is necessary in order to produce synthetic speech that is not limited to a speech corpus. A parametric method of concatenative synthesis is necessary to improve the quality of synthetic speech.

Harmonic plus Noise Model (HNM) has shown the capability of providing high-quality synthesis and prosodic modifications [6]. The result of the listening test conducted in [7] shows that HNM scores higher than TD-PSOLA in intelligibility, naturalness, and pleasantness.

This paper presents the incorporation of prosodic models of the duration and intonation of Filipino speech, and the application of HNM to Filipino speech synthesis. The next part of this paper is devoted to processes involved in building the prosodic models, and a description of the analysis and synthesis of speech based on HNM. This is followed by a presentation of the results from listening tests and a conclusion on the quality of the synthesized speech. Finally, recommendations on improving the quality of the synthesized speech are presented.

This work was supported by the Office of the Vice Chancellor for Research and Development, University of the Philippines Diliman, under the Open Grant.

II. Methodology

A. Data

The data were taken from two chosen speakers, one female and one male, in the Filipino Speech Corpus. The Paragraphs-Sentences (ParSen) sub-corpus and the Words sub-corpus were used. Both sub-corpora were transcribed at the phone level based on the DSP26 phone set and some additional phones, as shown in Table I.

TABLE I
MODIFIED DSP26 PHONE SET

/a/	/b/	/h/	/ng/	/t/	/z/
/e/	/k/	/j/	/p/	/ts/	
/i/	/d/	/l/	/r/	/v/	/epi/
/o/	/f/	/m/	/s/	/w/	/q/
/u/	/g/	/n/	/sh/	/y/	/pau/

The phones that were added were /pau/ (inter-sentence pauses), /q/ (glottal stop), and /epi/ (epinthetic silence or inter-phone silences).

Two corpora for each speaker, male and female, were prepared – a training corpus was used to create the models for the duration and intonation, and a synthesis corpus for the waveform synthesizer.

The synthesis corpus is a database of diphones extracted from the Words sub-corpora. A unique sample of every diphone in the middle of an isolated word and those that are in the boundaries were stored. Although it is recommended that diphones be extracted in the middle of an isolated word [8], only a few diphone samples were collected by following this recommendation. Diphones that are not in the middle of the isolated word were also stored.

The ParSen sub-corpus consisting of 81 phrases for female speaker and 96 phrases for the male speaker was also transcribed into syllables, words, and phrases. The pitch contour of the ParSen sub-corpora was extracted using an autocorrelation method.

B. Natural Language Processing (NLP) Module

The NLP module provides the necessary string of speech units and the correct prosody that was used by the Digital Signal Processing module in producing the synthesized speech. The speech units were extracted from the input text using a simple text analyzer. The text analyzer utilized a lexicon which translated the words in the input sentence to their corresponding phonetic spelling and then to diphones. For words that were not present in the lexicon, letter-to-sound rules were applied.

Duration Modeling

To produce the desired speed of the synthesized speech, a model of the duration of Filipino was developed.

A Classification and Regression Tree (CART) was used as the model for the durations. The CART was trained for Z-scores, or the normalized duration of the phone, using features extracted from the text. These features were the current, next, and previous phone, the classification of the phone according to voicing, the relative position in the syllable, word, and sentence, and the number of phones, syllables, and words in the current syllable, word, and sentence, respectively. Training was done using Wagon, the CART generation program of the Edinburgh Speech Tools suite using 80% of the ParSen subcorpus; i.e., 65 phrases for female speaker and 76 phrases for male speaker. The remaining 20% was set aside for the test set; 16 phrases for female speaker and 20 phrases for male speaker.

Intonation Modeling

To model the intonation of Filipino Speech, Tilt model was used. Tilt expresses the overall shape of the event. The intonation event label of the phone was added in the features used in the duration modeling. These features were used in the CART training for the Tilt parameters such as duration, amplitude, tilt, position of the peak, and starting pitch of the event. The duration is the sum of the rise and fall durations of the event. Amplitude is the sum of the magnitudes of the rise and fall amplitudes. Position of the peak is where the rising event stops and the fall begins. Starting pitch is the pitch at the point from which all other calculations may be made. The training was done using Wagon and the training set consisted of the same set used in duration modeling.

C. *Digital Signal Processing Module*

Harmonic Plus Noise Model (HNM) implemented in [9] was used to synthesize speech. The HNM has three phases: analysis, post-analysis, and synthesis.

HNM Analysis

HNM parameters were estimated in the analysis phase. An interval of the possible pitch values was defined to be between 40 to 400 Hz.

A speech segment was windowed with length that is dependent on the lowest expected fundamental frequency. A typical Hanning window was used.

Initial pitch was estimated by maximizing the equation

$$\Psi(P) = P \cdot \sum_{l=-\infty}^{\infty} r(l \cdot P) \qquad (1)$$

where P = period in the set $[fs/f_{omax}:fs/f_{omin}]$ and the function $r(k)$, which is an autocorrelation function, is defined as

$$r(k) = \sum_{t=-\infty}^{\infty} s(t)w^2(t)s(t+k)w^2(t+k) \qquad (2)$$

where s = speech signal

w = analysis window (Hanning window)

The initial pitch estimate was used to generate a synthesized signal $\hat{s}(t)$. The voiced/unvoiced decision was made by comparing the normalized error over the first four harmonics of the estimated pitch to a given threshold, -15 dB, as shown in (3).

$$E = \frac{\int_{0.7f_o}^{4.3f_o} (|S(f)| - |\hat{S}(f)|)^2}{\int_{0.7f_o}^{4.3f_o} |S(f)|^2} \qquad (3)$$

where $|S(f)|$= original spectrum and $|\hat{S}(f)|$ = synthetic spectrum.

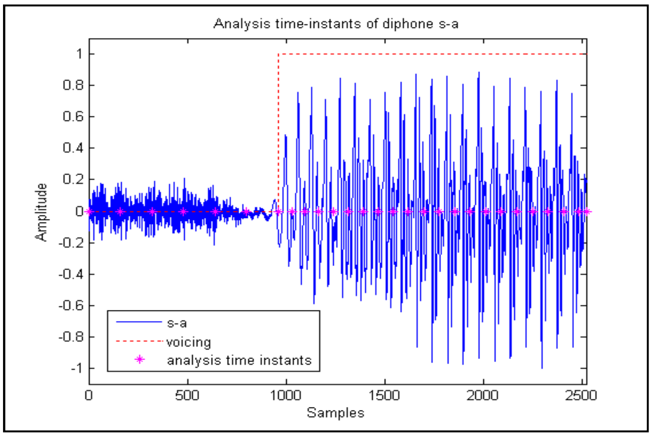

Fig. 1. Voicing decision of the diphone /s-a/ and Analysis time-instants placement.

An example of voicing decision of the diphone /s-a/ is shown in Fig. 1. The dotted line shows that frames with the phone /s/ were classified as unvoiced (value of 0), while frames with the phone /a/ were classified as voiced (value of 1). Analysis time-instants are indicated by stars.

For voiced frames, voiced frequencies were determined by a peak-peaking algorithm and 'harmonic test' [6]. The last voiced frequency determined the maximum voiced frequency (F_{mv}) of the current frame.

The initial pitch estimate was refined using the frequencies that were classified as voiced and minimization of the error criterion, $E(\hat{f}_o)$, shown below:

$$E(\hat{f}_o) = \sum_{i=1}^{L_n} |f_i - i \cdot \hat{f}_o|^2 \qquad (4)$$

where $\hat{f}_0$ = refined pitch

f_i = voiced frequencies

L_n = number of voiced frequencies

Using the stream of refined pitch values, the analysis time-instants t_a^i were set at a pitch-synchronous rate on the voiced portions of speech, and at 10 ms on unvoiced segments.

$$t_a^{i+1} = t_a^i + P(t_a^i) \qquad (5)$$

where $P(t_a^i)$ is the pitch period at analysis time-instant.

Amplitudes a_k and phases φ_k were estimated for each t_a^i by computing the complex amplitude (A_k) of the *kth* exponential which is given by:

$$A_k = \frac{\sum_{t=t_a^i-N}^{t=t_a^i+N} w^2(t)s(t)e^{-j2\pi k f_o t}}{\sum_{t=t_a^i-N}^{t=t_a^i+N} w^2(t)} \qquad (6)$$

where s = original signal

w = weighting function

N = integer closest to the local pitch period

f_0 = fundamental frequency

Noise parameters were estimated around each t_a^i. The original spectral density function was modeled by a 15th order AutoRegressive (AR) filter using a correlation based method.

HNM Post-Analysis

The target duration from the duration model and original duration of the diphones were used to compute the time modification factor β. Phone boundaries within the diphones were computed to have the duration of the two phones since they have different durations. This way, the diphones have different values of β which were determined by dividing the target duration with the original duration.

The target pitch contour from the intonation model was divided by the original pitch contour estimated in the analysis phase to determine the pitch modification factor α.

In the analysis phase, the HNM analysis windows were placed in a pitch synchronous way regardless of the position of the glottal closure instants. This simplifies the analysis process but increases the complexity of the synthesis. Phase mismatch between frames from different acoustic units must be considered. To solve this problem, a method for the synchronization of signals based on the notion of center of gravity [9] was applied. The method states that if the estimated phases $\hat{\phi}(k\omega_0)$ from the complex amplitude (A_k) at the frequency samples $k\omega_0$ are corrected by

$$\hat{\theta}(k\omega_0) = \hat{\phi}(k\omega_0) - k\hat{\phi}(\omega_0) \tag{7}$$

then all the voiced frames will be synchronized around their center of gravity. Using (7), the estimated phases $\hat{\phi}(k\omega_0)$ are replaced with $\hat{\theta}(k\omega_0)$.

The A_k was then interpolated to obtain an envelope of the complex speech spectrum.

HNM Synthesis

Synthesis time instants (t_s^i) must be estimated initially to be used in the synthesis of harmonic and noise parts of the synthesized speech. Shown in Fig. 2 is an example of synthesis time instants estimation. For synthesis with no prosody, t_a^i will also be t_s^i. In case of prosody modifications, the synthesis time instants were derived from the target pitch contour and time modification factor. First, the target pitch contour of each phone in the diphone was resampled to their respective desired duration. Then the synthesis time axis was generated by applying (8) and (9), where f_s is the sampling frequency, T_i is the pitch period, P is the resampled pitch contour and s is the location of the synthesis time instant, starting at $s_1 = 1$:

$$T_i = f_s \div P(s_i) \tag{8}$$

$$s_{i+1} = [s_i + T_i] \tag{9}$$

Then a virtual time axis was derived from scaling of the analysis time axis to be of the same length as the synthesis time axis to map the analysis time axis to the synthesis time axis. The mapping is done by finding the nearest analysis time axis element to the virtual time axis element. (10) is used for the mapping.

$$\underset{j=1,2,\ldots,N}{min}\left(\left|a_j - s_i\right|\right) \tag{10}$$

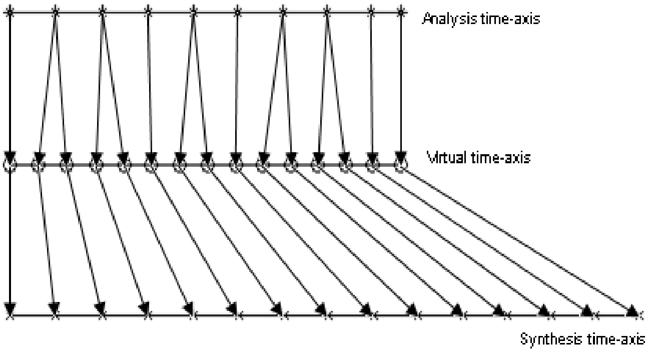

Fig. 2. Estimation of Synthesis Time-instants for Diphone /a-a/

For the synthesis of the harmonic part, the complex amplitude envelope was resampled at the harmonics of the fundamental frequencies (f_o) of every t_a^i for no prosody. In the case of prosody modifications, the fundamental frequencies were multiplied with the pitch scale modification factor α at every t_a^i to shift the harmonic frequencies. Then the complex amplitude envelope was sampled at the shifted harmonic frequencies.

For the synthesis of the harmonic part of a frame, (11) was applied.

$$h(t) = \sum_{k=-L}^{L} A_k(t_s^i) e^{j2\pi k f_o(t_s^i)(t-t_s^i)} \tag{11}$$

L is the number of harmonics included in the harmonic part: $L = F(t_s^i)/f_o(t_s^i)$ and $A_k(t_s^i)$ is the complex amplitude of the kth harmonic with $A_{-k} = A_k^*$.

For the synthesis of the noise part, a unit-variance white Gaussian noise was created and filtered using a normalized all-pole filter which was designed using the AR coefficients from the analysis phase. The filtered output was then multiplied with the envelope of variance. A high-pass filter with a cut-off frequency equal to the F_{mv} was created to filter the noise part of a voiced frame. The noise part was then synthesized by modulating the overlap-added frames with a time domain envelope which was synchronized with the pitch period.

The harmonic part and noise part were added to synthesize the diphone.

III. Testing and Results

To be able to measure the effectiveness of the duration model, intonation model and the synthesizer separately, eight phrases were synthesized for each test set phrase, with the original phrase as control.

The test contains 10 phrase sets, consisting of different phrase types, rated using the Mean Opinion Score (MOS) method, which is a scale of 1 (Bad) to 5 (Excellent). The 10 phrase sets were repeated three times and the median of the repetitions was taken as the score of the sentence.

Listening tests were conducted for the female and male speaker. For both speakers, the test was conducted with 15 listeners.

Regardless of the speaker, results (presented in Table II) show that the synthesized speech that uses duration and intonation models give output comparable to synthesized speech that uses actual prosody. HNM with prosodic models has similar performance to TD-PSOLA. HNM performs slightly better than TD-PSOLA when used with actual prosody. This shows that, with a better prosodic model, HNM can possibly score higher than TD-PSOLA.

The low scores obtained by simple concatenation of diphones shows the importance of prosody on synthesized speech.

Diphone samples are also major contributor to the quality of the synthesized speech. Some samples are too short such that voicing decisions tend to be erroneous.

TABLE II
LISTENING TEST RESULTS

Phrase Type	Female Speaker		Male Speaker	
	MOS	S.D.[a]	MOS	S.D.[a]
Original	4.93	0.26	4.90	0.29
HNM + Predicted Duration + Actual Pitch	2.80	0.59	2.72	0.58
HNM + Actual Duration + Predicted Pitch	2.89	0.66	2.63	0.53
HNM + Predicted Duration + Predicted Pitch	2.75	0.52	2.62	0.45
HNM + Actual Duration + Actual Pitch	3.00	0.62	2.73	0.55
TD_PSOLA + Predicted Duration + Predicted Pitch	2.75	0.60	2.63	0.54
TD_PSOLA + Actual Duration + Actual Pitch	2.83	0.55	2.69	0.53
Simple concatenation of diphones	2.35	0.87	2.43	0.73

[a] S.D. (standard deviation)

IV. CONCLUSIONS

The results showed that diphone concatenation using the HNM produces synthesized speech that is between poor (MOS of 2) and fair (MOS of 3) quality. This means that listeners tend to exert considerable to moderate effort in understanding the synthesized speech.

With a small number of training data (65 phrases for the female speaker, and 76 phrases for the male speaker), the results of the listening tests showed that the prosodic models of duration and intonation developed for the Filipino language were able to provide synthetic speech that requires considerable to moderate effort to be understood, e.g. MOS of 2.75 for female speaker & 2.62 for male speaker in the synthetic phrase that uses predicted prosody.

Using actual prosody, the score of the developed system that uses HNM is slightly higher than the score of the system that uses TD-PSOLA.

V. RECOMMENDATIONS

In order to improve the synthesis, we recommend the following.

1. Design and develop a corpus for speech synthesis. The corpus should contain more paragraphs and sentences to have more data with which to train the CART and more isolated words which will be able to cover all the possible diphones in Filipino language. In theory, increasing the training data would generate better models.
2. A corpus with nonsense words should be designed and developed. This will allow for a more systematic way of collecting unit speech samples or diphones. The diphones will be extracted in the middle of the non-sense words.
3. In modeling the prosody, punctuation marks should be added as features.
4. The transcriptions of the recordings must be as accurate as possible to have good samples of diphones for the database.

REFERENCES

[1] M. T. Corpus & J. J. Liampo (2001), Continuous Text-to-Speech Synthesis of Tagalog Words, Undergraduate Student Project, UP Diliman.

[2] M. Co (2002), “Prosody Development for Filipino Text-to-Speech Systems,” MS Thesis, College of Engineering, UP Diliman.

[3] A. Mendigorin, Jr. (2007), “Automatic English Text to Filipino Speech Translator Using Dictionary Based Translation and Concatenative Synthesizer with Prosody for Filipino Speech,” Undergraduate Student Project, College of Engineering, UP Diliman.

[4] L. Tupas (2002), “Concatenative Synthesis of Two-Syllable Filipino Words,” Undergraduate Student Project, College of Engineering, UP Diliman.

[5] J. Asis (2005), “Automatic Duration Modeling and Time-Scale Modification of Filipino Speech,” Undergraduate Student Project, College of Engineering, UP Diliman.

[6] Y. Stylianou (2001), “Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis,” *in IEEE Transactions on Speech and Audio Processing, Vol. 9, No. 1.*

[7] A. Syrdal, et.al. (1998), “TD-PSOLA versus Harmonic Plus Noise Model in Diphone Based Speech Synthesis,” *in Proc. of the International Conf. on Acoustics, Speech, and Signal Processing.*

[8] K. Lenzo & A. Black (2000), “Diphone Collection and Synthesis,” *in Proceedings of the International Conf. on Spoken Language Processing*. 1987, pp. 740–741 [*Dig. 9th Annu. Conf. Magnetics* Japan, 1982, p. 301].

[9] V. Vasilopoulos, A. Prayati, and A. Athanasopoulos (2007), “Implementation and evaluation of a Greek Text to Speech System based on an Harmonic plus Noise Model,” in IEEE Transactions in Consumer Electronics, pp 585-592.