# DEVELOPMENT OF A FILIPINO SPEECH CORPUS

Rowena Cristina Guevara, Melvin Co, Evan Espina, Ian Dexter Garcia, Emerson Tan, Ryan Ensomo, Ramil Sagum

DSP Laboratory, Department of Electrical and Electronics Engineering
University of the Philippines, Diliman, Quezon City, Philippines
dsp@eee.upd.edu.ph

**ABSTRACT**

In this paper we describe the Filipino Speech Corpus that we are developing. Read text and spontaneous speech will be recorded for 100 speakers. The text consists of paragraphs, sentences, words, syllables and phonemes. The spontaneous speech consists of a description of a movie that the speaker recently watched or an event that s/he recently experienced. The data is recorded at 44.1 kHz with each recording session lasting 20 to 72 minutes and each speaker attending 2 recording sessions. The data, re-sampled at 16 kHz, and the word level transcription for the read text will be available on April 5, 2003. The transcription for the spontaneous speech will be available six months later. The corpus development is a joint effort between the Linguistics Department and the Electrical and Electronics Engineering Department at the University of the Philippines, Diliman. The last two sections of this paper describe current projects that use the Filipino Speech Corpus and other projects that are lined up for Filipino speech research at the DSP Laboratory.

## 1. INTRODUCTION

Linguists and engineers have respectively pursued language and speech research in the Philippines separately. In this effort, the two groups have teamed up to develop a Filipino speech corpus with the following objectives: (1) provide data to stimulate research involving Filipino speech and contribute to the data used by language researchers, (2) gather prosodic information for Filipino speech synthesis, and (3) collect clean training data for Filipino speech recognition.

In the past three years, projects at the DSP Laboratory involving Filipino speech are isolated word recognition, synthesis-from-phones, speaker transformation and HMM-based codecs. In these projects, data was generated as needed. Now that we have the facility to record clean speech, we decided to design a corpus that will address the data needs of our speech research program. By making the corpus available on demand, we expect to encourage other researchers and applications development to use Filipino speech in their endeavor.

## 2. FILIPINO LANGUAGE AND SPEECH

There are more than 100 indigenous languages in the Philippines, and some of these languages have dialects [1]. Filipino is the national language and it is based primarily on Tagalog that is linguistically classified as an Austronesian or Malayo-Polynesian language.

The Tagalog alphabet consists of the following:

a b k d e g h i l m n ng o p r s t u w y

In the Filipino alphabet, the following have been added:

f, j, v, z

The Filipino language is spoken as it is spelled. For example, two successive vowels, such as the word ‘oo’ (English meaning: ‘yes’) will be pronounced as 2 syllables, ‘o-o’.

The Filipino language has many non-Tagalog words hewn from Spanish, Chinese, English and the other Philippine languages. These words are either re-spelled to reflect the pronunciation or mutated as word combinations [2]. Here are some examples:

Re-spelled:
Character (English) $\Rightarrow$ Karakter (Filipino)
Hypothesis (English) $\Rightarrow$ Haypotesis (Filipino)

Mutated:
Como esta (Spanish) $\Rightarrow$ Kamusta (Filipino)
Framework (English) $\Rightarrow$ Preymwork (Filipino)

## 3. CORPUS INFORMATION

Filipino speakers are recruited to volunteer their time in 2 recording sessions. There will be 50 male and 50 female speakers, ages 16 and older. The speakers are undergraduate students, lecturers or faculty members at the University of the Philippines, Diliman.

The speakers come from all over the Philippines and may have Tagalog as their first, second or third language. Each volunteer must sign a written consent form, whereby s/he agrees to have the recording of his/her voice, released, together with information regarding his/her first, second and third language, languages spoken at home and his/her parents’ first language.

The speaker's identity will not be part of the released information.

### 3.1 Read text

A copy of the prepared text is given to the speaker just before the recording session. There are 5 sets of text material for the paragraphs and sentences/phrases. Each set is divided into two, such that the speaker reads half of the text material in each recording session. All the speakers will be reading the same set of words, syllables and phonemes. The read text was formulated based on the recommendation of the Linguistics Department to elicit the phones and prosodic cues that characterize Filipino speech.

Each paragraph contains 31 to 109 words and has the following themes: (1) introducing oneself; (2) relating an emotional event; (3) story-telling; (3) giving directions; (4) giving advice; and (5) describing a scene. Here is an example together with the English translation:

> Nagpapakilala: Magandang umaga. Ako si Lara. Ako ay 27 anyos. Wala pang asawa ngunit gusto ng magkapamilya. Sa ngayon ay wala pa akong boyfrend. Pero marami na rin ang nanliligaw sa akin.
>
> Introducing oneself: Good morning. My name is Lara. I am 27 years old. I am not married by I wish to have a family. Right now, I do not have a boyfriend. But I have many suitors.

Each sentence/phrase in the text consists of 1 to 7 words. The speaker will be reading 5 questions, 3 commands, 2 requests and 2 greetings/exclamation in each recording session. Here are some examples and their English translation:

> (Patanong) Saan pala ang libing?
> (Question) By the way, where's the funeral wake?
>
> (Pautos) Gumising ka na.
> (Command) Wake up.
>
> (Pakiusap) Pakibukas naman ang pinto.
> (Request) Please open the door.
>
> (Pagbati atPadamdam) Aray!
> (Greeting or Exclamation) Ouch!

Each speaker reads a total of 748 words in each recording session. The longest word has 19 characters: 'maghintay-hintay' (in English: wait for a while). The word with the most number of syllables: 'mapapakinabangan' (in English: can be of benefit), has 7 syllables.

A Filipino syllable can have 1 to 3 phones and can be a vowel: a, e, i, o, u; consonant-vowel: ba, be, bi, bo, bu, ka, ke, …, za, ze, zi, zo, zu; vowel-consonant: an, on, is; or consonant-vowel-consonant: ping, syem, tsis, vays. Only the syllables of the 2nd and 4th kind are recorded. Each speaker reads a total of 375 syllables in each recording session.

In each recording session, the speaker utters the following Filipino phonemes: a, b, d, e, f, g, h, i, j, k, l, m, n, ng, o, p, r, s, t, ts, u, v, w, y, z.

The speaker's spontaneous speech is recorded in the last part of the 2nd recording session. The speaker describes a movie that s/he recently watched or an event that s/he recently experienced. The spontaneous speech lasts no more than 5 minutes.

### 3.2 Recording specifications

The recording is done in an isolation booth that has heavily padded walls, ceiling and floor. A microphone connected to a digital-to-digital card through a digital audio tape samples the speech at 44.1 kHz, and the data is stored as a mono, 16 bit *.wav file. The microphone is wireless, cardioid and fixed on a boom stand. We measured a 50 dB SNR in the entire recording setup.

In a recording session, the speaker is given a copy of the text that s/he will read and instructed to read the text at normal speaking rate and pronunciation. Each recording session lasts 20 to 72 minutes, depending on the speaking rate of the speaker.

The data is mastered onto a CD for archiving, and down-sampled to 16 kHz for use in the laboratory.

### 3.3 File naming convention

The names of the sound files have the following information: speaker identification number, speaker gender, speaker age group and text material set. For example, the file name, 12-0-4-2.wav, refers to

12 --speaker id no.
0--gender
  0-male
  1-female
4-age group
  0:20-27
  1:28-35
  2:36-43
  3:44-51
  4:52-60
2-text material number

### 3.4 Transcription

Using an HMM trained on 2 classes, speech and non-speech, an initial pass is done on the waveform to segment it into these classes. The resulting segmentation is checked for early or late onsets and offsets, and for any missing segment breakpoints. The segmentation is corrected accordingly before it is passed on to the labeling phase.

In the labeling phase, the text that was used in the recording session is used as a basis for labeling each segment. For the read paragraphs, the segments are determined by the pauses of the speaker either at the punctuation marks, or to take a breath in the

middle of a long sentence. The read sentences/phrases and words are segmented on a sentence/phrase and word level, respectively. Ideally, the read syllables can be segmented on the syllabic level. However, some speakers utter the syllables such that co-articulation occurs between syllables. In such cases, the series of syllables are lumped in one segment.

A last pass is done to transcribe non-vocalic and vocalic noise, interrupted words and re-starts. Both non-vocalic and vocalic noise are labeled {NOISE} and a '-' is used to label an unfinished word or a re-start. We use the Transcriber® tool to generate an XML file, as shown in Figure 1.

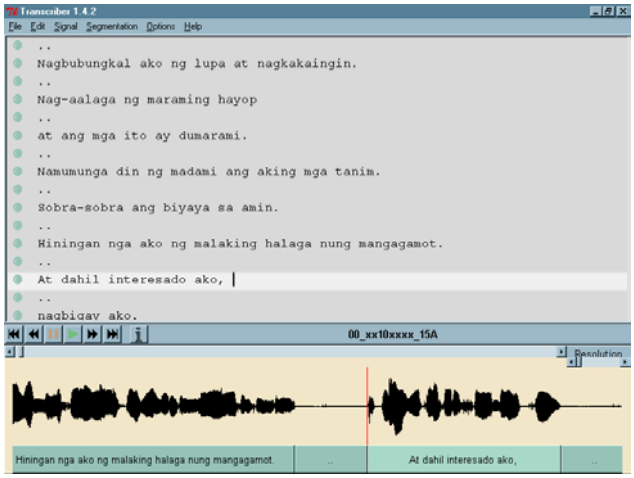


**Figure 1. Sample transcription using Transcriber®**

## 4. Practical applications

The data gathered so far, for 26 speakers, are/were used in several projects. These projects are described briefly below.

### 4.1 Real-time implementation of a low bit rate Filipino speech codec using Hidden Markov Model-based speech recognition/synthesis

A Texas Instrument® TMS320C6x–based codec is built with an encoder that extracts Mel-Frequency Cepstral Coefficients (MFCC) and pitch information from the input speech segment. The extracted MFCCs are the input to an HMM-based phoneme recognizer. Phoneme indices, state transitions and pitch information are then transmitted over an ideal channel. At the decoder, the corresponding MFCCs are generated based on the received phoneme indices and state transitions. The speech is synthesized using a Mel-Log Spectrum Approximation filter [3, 4]. The input to this filter is an excitation signal generated from the received pitch information.

The implemented speech coder achieved a bit rate of about 1.2 kbps, which is an order of magnitude below the conventional telecommunication rate of 64 kbps. The tradeoff with this low bit-rate is the loss of the identity of the speaker. For this task, the measure for success is the subjective intelligibility of the synthetic output. A subjective listening test was conducted to rate the synthetic speech based on two aspects: listening quality and listening effort. The latter refers to the effort require to understand the output. Using a grading scale based on the ITU-T recommendations, the codec netted a Mean Opinion Score (MOS) of 2.394 and 2.321 for listening quality and listening effort, respectively. On the ITU-T scale, a score between 2-3 is considered intelligible.

### 4.2 Prosody development for Filipino TTS system

In this project, an automated prosody overlay for Filipino Text-to-Speech System will be developed using the corpus [5]. A suitable segmental unit, such as syllable, phoneme or diphone, will be determined from the corpus. The speech corpus will then be segmented into these units and the F0 contours of each segment will be extracted. The set of all possible intonation phrases that exists in Filipino speech will be determined from the collection of extracted F0 contours [6]. A local maxima and minima of the F0 contours will be set prior to deciding the number of tones that are needed to represent Filipino speech. The existence of downdrift effect will be taken into account and observed for consistency. The rate of declination in Filipino speech will also be determined.

Based on the set of intonation phrase, a phonetic model will be developed. This model will serve as a mapping from the phonology level to F0 level. The F0 contour from the F0 level, together with the chain of sound elements, will be the input to the speech synthesizer. The synthesizer will be implemented using either Sinusoidal Modeling [7] or Harmonic Plus Noise Modeling [8].

### 4.3 TIMIT-Bootstrapped Filipino phoneme recognizer

A Filipino phoneme recognizer was built by bootstrapping from a phoneme recognizer trained on TIMIT. The recognizer is applied to the uttered syllables in the FIlipino corpus.

Cepstral coefficients with deltas and double deltas were computed for 3676 TIMIT sentences. The normalized features were used to train an MLP with 500 hidden units; 10% of the training sentences were used for cross-validation. The training sentences are as specified in the TIMIT documentation, and the target was the TIMIT 61 phone set. A stack decoder [9] that uses posterior probability estimates, an HMM-based phone model, and language model was used to test the recognizer on 1339 TIMIT test sentences. The average phoneme recognition rate is 70.3%, where the error is defined as phoneme substitution, deletion and insertion.

When the recognizer was applied to Filipino syllables, the average phoneme recognition rate was 32.08%. An analysis of errors shows that the language model was the primary cause of error. When a Filipino phonotactic model was used as an input to the stack decoder, the average phoneme recognition rate went up to 42.12%. The recognition results for each Filipino phoneme are shown in Figure 2.

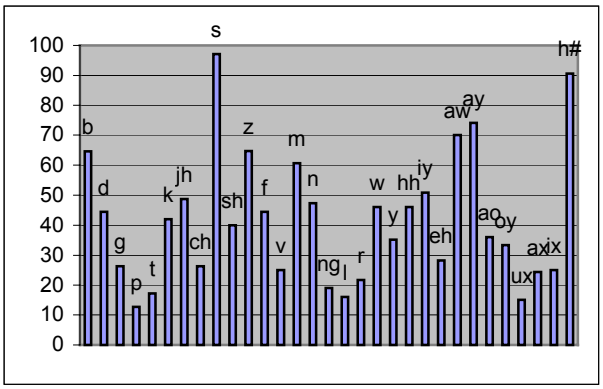


**Figure 2. Recognition rate for each Filipino phoneme**

### 4.4 Automatic Filipino phoneme recognition using MLP and a start-synchronous decoder

The best Filipino phoneme recognition rate from bootstrapping the recognizer from TIMIT is very low to be used as a basis for word-level or continuous speech recognition. There is a need to phonemically transcribe the Filipino Speech Corpus in order to improve the phone model, and correspondingly improve the phoneme recognition rate. Our approach is to iteratively increase the size of the training set for the recognizer and use the results to phonemically segment and recognize another set of recording. The initial results for a training set with a size of 2666 and 590 hidden units are 81%.

## 5. Conclusion

This paper discussed the development of a Filipino Speech Corpus and the projects that have started using the available data. As we develop the corpus, a language model is being developed based on Filipino news articles on the Internet [10].

The word-level transcription of the Filipino Speech Corpus can be used to develop a word-level speech recognizer. The phoneme-level transcription of the corpus can be used to develop a continuous speech recognizer. The corpus may also be used for multi-unit speech synthesis.

The other applications that may avail of the corpus are phonologic, phonetic and linguistic research, human-machine communications and multi-lingual speech analysis.

We expect the corpus to be around 12 GB and therefore the data can be made available as a set of 18 CDs or one removable hard disk.

## 6. Acknowledgements

The authors would like to thank the International Computer Science Institute for the use of their software and the help extended by members of the Speech Group, the University of the Philippines System and the Banatao Fellowship for supporting this effort.

## 7. References

[1] E.A. Constantino, "Current Topics in Philippine Linguistics"', *Meeting of the Linguistic Society Of Japan,* Yamaguchi, 1998.

[2] J.G.U. Rubrico, "The Metamorphosis of FIlipino as National Language", http://www.seasite.niu.edu/Tagalog/essays_on_philippine_languages.htm

[3] T. Fukada, T. Kobayashi and S. Imai, "Speech Parameter Generation from HMM Using Dynamic Features," *Proc. ICASSP-95, pp. 660-663,* 1995.

[4] T. Fukada, T. Kobayashi and S. Imai, "An Adaptive Algorithm for Mel-Cepstral Analysis of Speech," in *Proc. ICASSP-92, pp. 137-140,* 1992.

[5] B. Mobius, "Corpus-Based Speech Synthesis: Methods and Challenges," Institute of Natural Language Processing, University of Stuttgart.

[6] K.N. Ross, and M. Ostendorf., "A Dynamical System Model for Generating Fundamental Frequency for Speech Synthesis," *IEEE Transactions on Speech and Audio Processing,* pp 295-309, Vol 7, No. 3, May 1999.

[7] G. Bailly, E. Bernard, and P. Coisnon, "Sinusoidal Modelling," Institut de la Communication Parlee-INPG & Universite Stendhal, October 16,1998.

[8] Y. Stylianou, "Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis," *IEEE Transactions on Speech and Audio Processing*, pp 21-29*,* Vol. 9, No.1, January 2001.

[9] S. Renals, and M. Hochberg, "Start-synchronous search for large vocabulary continuous speech recognition. *IEEE Trans. Speech and Audio Processing*, 542-553, Vol. 7, 1999.

[10] E. dela Vega, *Language Model for Filipino*, Digital Signal Processing Laboratory, University of the Philippines, 2002.