PAPER

Open Domain Continuous Filipino Speech Recognition: Challenges and Baseline Experiments

Federico ANG[†*a)], **Rowena Cristina GUEVARA**[†], ***Nonmembers***, **Yoshikazu MIYANAGA**[††], ***Fellow***, **Rhandley CAJOTE**[†], ***Member***, **Joel ILAO**[†**], **Michael Gringo Angelo BAYONA**[†], ***and*** **Ann Franchesca LAGUNA**[†], ***Nonmembers***

SUMMARY In this paper, a new database suitable for HMM-based automatic Filipino speech recognition is described for the purpose of training a domain-independent, large-vocabulary continuous speech recognition system. Although it is known that high-performance speech recognition systems depend on a superior speech database used in the training stage, due to the lack of such an appropriate database, previous reports on Filipino speech recognition had to contend with serious data sparsity issues. In this paper we alleviate such sparsity through appropriate data analysis that makes the evaluation results more reliable. The best system is identified through its low word-error rate to a cross-validation set containing almost three hours of unknown speech data. Language-dependent problems are discussed, and their impact on accuracy was analyzed. The approach is currently data driven, however it serves as a competent baseline model for succeeding future developments.

key words: *filipino speech, automatic speech recognition, statistical learning*

1. Introduction

Continuous automatic speech recognition (ASR) systems are being used in a variety of software and devices. These ASR systems are assured of achieving high performance even under noisy circumstances. In order to develop such systems, a speech training database is mandatory. It is well known that the size and quality of the speech database is as important as the algorithms used for improving the ASR system performance. In other words, the database makes a large contribution to the baseline performance of the ASR system.

In this paper, a general-purpose ASR system is developed for the Filipino language. We provide baseline models for comparing various techniques applied to recognizing continuous Filipino speech in different domains. These results, in turn, can be used as a survey for the best working parameters and conditions when the need to develop task-specific ASRs arises.

Due to the variety of speech recognition tasks, under certain conditions the major processes of training, development, and evaluation oftentimes become more data-driven. For example, in our evaluations we did not consider an explicit analysis of the channel conditions of the ASR environment. Thus, there is a need to clarify the parameters and the dimensions of the system. The results and discussion presented in this paper are refinements of the work presented in similar reports on this topic. As with other papers on Filipino ASR [1]–[3], this report is by no means exhaustive and only serves to provide details about more recent developments. However, the evaluation presented in this paper is the first of its kind, due to the inclusion of loan words and informal syntax of the natural language. These characteristics are essential in reflecting the practical nature of the language, which also reflects the usefulness of the system. Also, other Filipino speech recognition developments mention data sparsity and bias issues on language modeling and evaluation [4]. Aside from this, there are other challenges in Filipino ASR that must be made known to give research ideas for those who wish to contribute in this ongoing development.

The paper is outlined as follows: Section 2 gives a description of the characteristics of the Filipino language, and discusses some challenges faced in developing the ASR system. Section 3 introduces the creation of a new speech database for Filipino speech systems development. Section 4 describes the dimensions of our running speech recognition system development. Section 5 gives a discussion of the results of the initial development stages of our ASR system. Section 6 focuses on the analysis and overall view of the problems in recognizing Filipino utterances. Section 7 gives some details regarding the use of our system for captioning television news broadcasts in Filipino under general conditions. Finally, Sect. 8 gives a summary of the results and provides recommendations for future developments.

2. Practical Considerations in Filipino ASR

In this section we provide specific information about the Filipino language, challenges in development, and data acquisition. While several details are mentioned in this paper, due to time constraints, the experiments done are still mostly data-driven and there are obvious details that we have yet to exploit.

2.1 Defining Filipino

Filipino is the national language of the Philippines, and is

Manuscript received December 17, 2013.
Manuscript revised April 24, 2014.

[†]The authors are with the DSP Laboratory, University of the Philippines, Diliman, Quezon City, 1101 Philippines.
[††]The author is with the ICN Laboratory, Hokkaido University, Sapporo-shi, 060–0814 Japan.
[*]Presently, with the ICN Laboratory, Hokkaido University.
[**]Presently, with the Computer Technology Department, De La Salle University, Manila 1004, Philippines.
a) E-mail: see http://csw.ist.hokudai.ac.jp/~ang

DOI: 10.1587/transinf.2013EDP7442

Table 1 Combined inventories [6], [7] of consonant sounds in Filipino (based on IPA symbols).

	Labial			Dorsal		Coronal			Laryngeal
	Bilabial	Labiodental	Labial-velar	Palatal	Velar	Dental	Alveolar	Post alveolar	Glottal
Plosive	p b				k g		t	d	ʔ
Nasal	m			ɲ	ŋ			n	
Trill								r	
Tap/Flap								ɾ	
Fricative		f v			x	θ ð	s z	ʃ ʒ	h
Affricate						t͡s		t͡ʃ d͡ʒ	
Approximant			w	j					
Lateral approx.								l	

spoken by 65 out of the 76 million Filipinos according to the 2000 national census. Philippine legislators have been careful in introducing provisions for the language as there is no mention of Filipino being based on the former national language Tagalog, whose name was later changed to Pilipino (and now, Filipino) to avoid association to any particular language, among the more than 100 Philippine languages. However, linguistic rules show that Tagalog and Filipino share identical grammar [5].

Because of the socio-political issue of resolving whether Filipino should be a single indigenous language based on the most influential group, or a mix of numerous indigenous languages, Filipino is still in the process of standardization and intellectualization. Thus, we are forced to rely on statistical information and base the description of Filipino on what is in widespread use. Due to this situation, some interesting challenges become present in developing its ASR system.

2.2 Filipino Language Characteristics

Here we mention some characteristics of the Filipino language, which present the actual challenges in the development of its continuous ASR system in both acoustic and language modeling aspects.

2.2.1 Phonology and Phonotactics

Combined inventories [6], [7] that accommodate loan words tell us that Filipino can have a maximum of about 45 phonemes, 29 of which are consonants and the rest are vowels. This maximum is based on the possibility of including users who are near native speakers of English. Table 1 shows the combined inventory of consonants. As for vowels, the counts come from the five basic vowel sounds of /*a*/, /*e*/, /*i*/, /*o*/, and /*u*/, four diphthongs /*aw*/, /*iw*/, /*aj*/, and /*uj*/, and the allophones from English.

2.2.2 Borrowing and Code-Switching

For a truly practical Filipino ASR, an important detail that we have to consider is that Filipinos use English in their daily conversations [8]. This code-switching can happen either within a sentence or as a complete sentence. For example, the sentence "Where *ba tayo magla*-lunch or something?" (Where are we going to have lunch or something?) has both Tagalog, English, and an English noun (lunch) applied with Tagalog morphology. Tagalog, by itself, has a very strong grapheme to phoneme relationship but this is perturbed by the presence of commonly used loan words not only from English, but also from Spanish, and other languages. Numbers and times can be spoken in English or in Spanish, or even in the native depending on the situation. Aside from these, differing accents in English complicates things further. Clearly, these characteristics make it harder to produce suitable acoustic and language models.

2.2.3 Complex Morphology

Contractions and the agglutinative morphology of Filipino have a big effect on the lexical coverage of the language model to be created. This morphology even provides a means of creating verbs from nouns, which can add to the richness of the vocabulary. For example, the English noun *internet* can be turned into a present progressive verb *nag-iinternet* (using the internet) by using Tagalog affixation and morphological rules. Intuitively, taking advantage of common affixation rules (e.g. morpheme strategies) can remedy this by giving them special attention in the language model.

2.2.4 Syntax

Different situations call for different syntax in Filipino. Informal spoken Filipino tends to be very flexible, while poliler forms give way to plural inflections and the insertion of words like *po* and *opo*, with *po* technically can be placed on different positions in a Filipino sentence. For example, one can casually say *nasaan ka?* (where are you?) but it is also possible to insert *po* to sound polite: *nasaan ka po?* or when in a different position makes an inflection, *nasaan po kayo?* The structure of sentences used by diplomats in public speaking or those in broadcasting also have their own characteristics as they use certain jargons that are not in casual sentences.

2.2.5 Non-Standard Orthography

Despite the existence of several orthographic rules coming from the commission appointed to be the arbiter of the national language, every existing written document in Filipino

cannot be regarded as already highly consistent in use or spelling. This affects consistency in transcriptions and adds confusability in the recognition. For example, the word for few can be spelled either *konti* or *kaunti* with the former spelling stemming from the fact that this is now the casual pronunciation for the word. These are solved by post-filtering the hypotheses using a list of words with the same connotation before the evaluation of alignments after decoding.

3. Tagalog Speech Corpus

The Digital Signal Processing Lab of the University of the Philippines, Diliman, is currently conducting a speech corpus development project under the ISIP Program funded by the Philippine government's Department of Science and Technology. This project will cover the 10 most widely spoken languages in the Philippines. Among the 10, Tagalog is one of those completed as of this writing. This is the first recognition system to use this database containing 100 hours of Tagalog speech. Being relatively new, licensing terms and distribution conditions are still being worked out. Relevant details can be inquired from one of the authors as they become available.

3.1 Recording Set-Up

Recording software was used to facilitate the data collection process, displaying the text to be read one sentence at a time. Every recording session has a guide to ensure quality of the recordings. All speech data were recorded at a high rate via a TASCAM DV-RA1000, then digitized and down-sampled to 16-kHz, at 16-bit resolution per sample, mono channel. The recordings were made using two set-ups, depending on whether the activity was done in the lab or off-site. Recordings inside the laboratory were done inside a pseudo-anechoic chamber using an AKG C414 XLS recording microphone interfaced with a Desktop PC via a Behringer XENYX2222FX Mixer and an M-Audio Mobile Pre-amp. The mixer was used to filter noise and allow two-way communication between the speaker and the recording personnel while avoiding mixing the audio signals in the same channel. The M-Audio MobilePre was used to connect the feed from the mixer to the Desktop PC. A visualization of this setup is shown in Fig. 1. The off-site recording required a much simpler set-up, involving only a Sennheiser CC 530 headset directly connected to a laptop.

3.2 Database Contents and Transcriptions

Approximately 300 prompts that contained isolated words, phrases, and sentences are generated as reading materials for every recording session. These prompts are automatically associated as the transcripts of the generated recording file. Because of the data-driven nature of our experiments, there was no consideration for a phonetically balanced prompt selection. The prompts were taken from a variety of sources

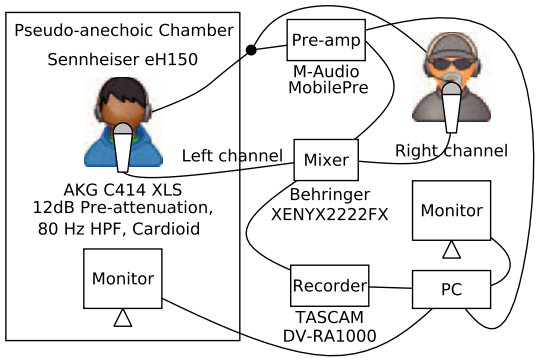

Fig. 1 On-site recording setup.

Table 2 Tagalog speakers demographics.

Location	Male	Female	Location	Male	Female
BATAAN	4	4	MINDORO	8	4
BATANGAS	9	12	NUEVA ECIJA	7	12
BULACAN	15	13	PALAWAN	3	3
CAM. NORTE	2	2	QUEZON	8	8
CAVITE	13	17	RIZAL	13	14
LAGUNA	10	12	TARLAC	2	2
MARINDUQUE	2	2			

with different domains, ranging from news, literary works, daily expressions, and situational conversations. These were either crawled from the internet or taken with permission from digital publications. Spontaneous speech were also recorded at the end of each session where the speaker is asked about several random topics. The transcriptions of these spontaneous speech were manually made with the help of native Tagalog speakers. Each utterance recorded is limited to a span of one minute.

3.3 Speaker Distribution

Variability in demographics has a huge impact in speech data collection quality and efforts were made to meet this feature. As a guideline, we identified 13 locations from the Philippine census and defined the number of participants that will be proportional to the number of Tagalog speakers in the 13 locations. Table 2 shows the result of this consideration. For Tagalog, a total of 201 native speakers participated in the recording sessions.

4. Speech Recognition System

4.1 Speech Data

Because the speech corpus and the recognition system were simultaneously being developed, only a subset of the speech corpus described was used. This subset contains 42 male and 62 female speakers. The age range of this subset is from 15 to 60 years old, with an average of 22.5 years. The average duration of a single utterance is about six seconds, or 11.4 words in average per utterance.

To compensate for moderately noisy environments,

Table 3 Filipino speech data statistics.

	Train	Dev	Eval	Total
Speakers (unique)	146 (144)	5	5	156 (154)
Utterances	33,340	1,525	1,527	36,392
Prompts covered	12,474	1,498	1,509	15,481
Words	375,039	19,961	19,100	414,100
Vocabulary	14,461	4,384	4,163	15,673
Duration in hours	54.9	2.8	2.8	60.6

such as in coffee shops or in a room with a few people, we included additional speech data containing such channel conditions. These new data consisted of 25 male and 25 female speakers. These data were also collected due to a large number of overlapping prompts in the original set. A summary of the resulting division for the training, testing, and evaluation sets is given in Table 3. Note that the vocabulary counts will exceed the total when added up due to overlaps.

4.2 Lexical Data

4.2.1 Phoneme Set

Originally, we planned on using the phonemes used in the CMU Pronouncing Dictionary [9] to fill in the richness of vowel sounds in English. However, by random sampling of the speech data, we observed that most of the English speech were not really adhering to most pronunciations found in the CMU dictionary. A previous study conducted on English-Filipino vowel mapping [10] suggests that there is really a close correlation between Filipino vowels and the actual sounds most Filipinos make when they speak English. Based on this notion, we settled on the following set (using IPA symbols): a, ɑ, ɐ, b, t͡ʃ, d, ð, e, ɜ, f, g, h, i, d͡ʒ, k, l, m, n, ŋ, ɲ, o, p, r, s, ʃ, t, θ, u, v, w, l̩, j, and z. The syllabic consonant l̩ (e.g. sing**le**) was considered due to its frequency in the corpora. The ɜ was later decided to always have an inherent r sound at the end due to its only use in the lexicon.

4.2.2 Pronunciation Lexicon

To expedite the creation of the pronunciation lexicon, the usual grapheme-to-phoneme technique was used. For Filipino, the rules are very simple and most entries are easily constructed automatically. Loan words however, which constitutes about 20% of the total vocabulary, are more dependent on manual checking of how the speakers pronounced each word and had to undergo several revisions. This lexicon contains about 16k unique words, 300 of which have pronunciation variants. Out-of-vocabulary (OOV) words are not reflected on testing as all words in the test set are found in the lexicon.

4.3 Data Organization

As mentioned in the details of gathering the speech data, overlaps in utterances existed across all speakers in the database. This is because of the automatic selection of prompts from the initially limited text database. Additional prompts were added to the database yet overlaps were still inevitable, though at a much lower scale. From this, it was decided to use all speakers from the first set only for training. The additional set then became the source of testing. Based on the average length of produced speech recordings from a single speaker, it was decided to have five speakers each or about two hours each of speech data for development and evaluation. The rest are included in the first set for training.

To determine the best sets, we had to minimize the bias. A ranking was done to the 50 speakers of the additional set based on several factors such as percentage of overlaps in utterances, total duration, total utterances, and so on. The top 10 was divided equally in terms of rank and gender for both the development and evaluation sets.

4.4 Feature Extraction

Feature extraction for the speech data was based on the commonly used front-end for our tool (see Sect. 5). This front-end makes use of Mel-Frequency Cepstral Coefficients (MFCC). Features from 16 ms. frames are extracted every 10 ms., generating 16 coefficients from a Mel-filterbank of 30 filters. Per-utterance-based cepstral mean and variance normalization are then applied. The computation of derivatives for temporal dependency is done via a linear transformation using 15 frames around the current frame, generating a 240 dimensional super vector. Linear Discriminant Analysis (LDA) is then applied to bring down the dimension to 48 while still keeping the distinctive quality of the features. Vocal tract length normalization (VTLN) in the linear domain [11] is later applied in a per-speaker basis as an enhancement.

4.5 Acoustic Modeling

Speech sounds (i.e. phoneme set) were modeled using the three-state left-to-right HMM topology with self transitions. Silence and a couple of non-speech events (for better alignment) were modeled using the same three-state topology, but without self-transitions on the first two states. Only fully continuous systems are considered, with a limit of 2000 and 3000 maximum distributions and codebooks when context is considered. Since labels are used, the HMMs are recursively trained using the Viterbi algorithm with Maximum Likelihood (ML) as the training criterion. The distributions are later applied with recursive Gaussian splitting [12] with a maximum of 64 Gaussians, and later improved by a variant of semi-tied covariance (STC) [13] training called Optimal Feature Space (OFS) training. This latter training results to a global invariant transformation matrix [14] that incorporates the LDA matrix computed previously.

The initial model for our recognition system was created by bootstrapping a small English seed model and using Forward-Backward algorithm to force-align initial labels for

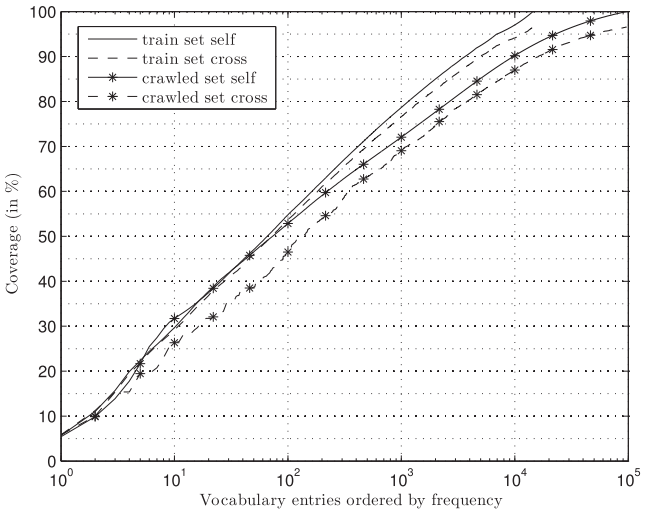

Fig. 2 Self- and cross-coverage plots of the training (train) set and the 13 Mio (via crawler) words set to the test set.

the training data. Successive training and label enhancing were made until degradations or plateaus in improvement were reached. We started with a context independent system, repeatedly trained it and later used it to bootstrap a quintphone-based context dependent system. There were 54 manually-created classes for the phone set for state tying.

4.6 Language Modeling

Language modeling (LM) and training were done using the Stanford Research Institute Language Modeling Toolkit (SRILM) [15]. Before actual modeling was done, some analysis were made via scripting.

4.6.1 Text Data

Aside from the sparsity of the training text due to the overlapping utterances across speakers, it was also observed that we cannot include the training text for language modeling because of its high correlation to the testing set utterances. We prepared subsets of texts from the remaining prompts not covered by the database then correlated the texts to the test corpus via a maximum likelihood-based weighted interpolation scheme. For all cases, the training corpus constitutes the mixture by 99%. Because of this, the training set was instead considered as the heldout set. After removing the overlaps with the testing set, it was used as the tuning set for LM training.

We then resorted to using three different text sets obtained from the internet through a crawler, which were available during the experiments. Figure 2 shows how the self- and cross-coverage of the training set are very close to each other while the texts from the crawler are more variable. Note that for the vocabulary of the corpora from the internet of 360k words, the OOV rate for the test set was still at 3.44%.

The vocabulary for the LMs was generated by taking the most frequent words from the corpora obtained via web crawler before a cutoff and crossing it with the words from the training set. This provided a 10.5k vocabulary, with an OOV of 5.45% when compared to the vocabulary from the test corpus.

Table 4 Language model perplexities on test set.

Smoothing	**Back-off**	**Interpolated**
Natural (Ristadt)	341.24	–
Good-Turing/Katz	272.52	–
Witten-Bell	268.48	369.14
Absolute (Ney)	267.75	375.48
Orig. Kneser-Ney	267.73	294.68
Mod. Kneser-Ney	**266.94**	279.66

4.6.2 LM Training

Using the three subcorpora crawled from the web, 4-gram back-off and interpolated models for several smoothing techniques [16]–[21] were generated. For each smoothing algorithm, the three LMs $P_1(w|h)$, $P_2(w|h)$, and $P_3(w|h)$ were combined to generate an interpolated model $P(w|h)$. This was done linearly via

$$P(w|h) = \lambda_1 P_1(w|h) + \lambda_2 P_2(w|h) + \lambda_3 P_3(w|h) \quad (1)$$

where interpolation weights λ_1, λ_2, and λ_3 were chosen to maximize the likelihood of the tuning set. After generating all models, we picked the one that minimizes the perplexity over the test set. Table 4 shows the results of the training. While differences are subtle for most cases, it is clear that the modified Kneser-Ney algorithm still gives the lowest perplexity. However, while most recognition systems get better results from interpolated n-gram models, our system got better perplexities from the simple backoff models.

4.7 Decoding and Speed

Lattice rescoring based on language model weight z and word transition penalty p variations is done to search for the necessary statistical correction for the combined acoustic and language model (log) scores:

$$P(W|\mathbf{X}) = P(\mathbf{X}|W)P(W)^z p^{|W|} \quad (2)$$

where $P(\mathbf{X}|W)$ is the acoustic model probability, $P(W)$ is the language model probability, and $|W|$ is the length of the utterance. The standard Word Error Rate (WER) is used to evaluate all systems.

Decoding parameters such as beam settings were heuristically determined from previous experiments that balances recognition accuracy and speed. Feature Space Adaptation (FSA) [22] was applied in decoding based on empirical results where it consistently gives significant improvements to our systems. This advantage, however, is expected to go down to about 1% as the system improves. The speed of our tool's single pass decoder [23] is at a real-time

factor average of 0.13 and 0.21 using the best decoding parameters, for context-independent and context- dependent systems respectively, using a 3.6GHz Intel Core i7-3820 computer. This fast decoding enables us to use the systems for real-time applications.

5. Evaluation Results

All experiments for the speech recognition system were done using the Janus Recognition Toolkit (JRTk) and the IBIS decoder [23], jointly developed at the Karlsruhe Institute of Technology (KIT) and the Interactive Systems Lab (ISL) at Carnegie Mellon University (CMU). Alignments and scores were generated using the Speech Recognition Scoring Toolkit (SCTK) [24] from the Information Technology Laboratory (ITL) of the National Institute of Standards and Technology (NIST).

5.1 Context Independent

The bootstrapped model was evaluated on the test set and initial results were at 51.5% and 58.6% WER, for the FSA-based and non-FSA decoding respectively. From successive evaluations it was observed that using FSA gives around 3% average advantage over its non-FSA counterpart. A set of recognition scores from the first labeling epoch are given in Fig. 3 for both development and evaluation sets. From this set of results, we decided to use seven Viterbi iterations for succeeding evaluations. In the next set of tables, only FSA-based results are shown. Table 5 summarizes the results of the successive label writing procedure done using the development training and test sets under a comparable set up of seven Viterbi iterations.

As can be observed, the best results were achieved from the second writing of labels and performance degrades after subsequent alignments. This behavior can be attributed to the large training data set. After further training, we achieved the lowest WER for the context independent system at 30.0%.

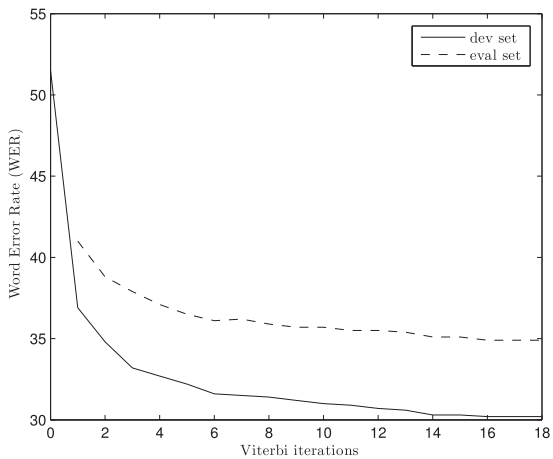

Fig. 3 Word Error Rates for the development and evaluation sets of the first written labels.

5.2 Context Dependent

A similar training procedure was done for the context dependent task, but this time, variation was made to the number of distributions. Table 6 shows the result for the experiments.

It is worth noting that this increase in distributions does not have any impact on the real time performance of the system mentioned previously. Again, the best performance was achieved on the second labeling epoch. After recursive training, the best performance on the development set was at a WER of 21.7%.

5.3 Baseline Enhancements

Table 7 shows the summary of the performance of the best trained systems and further enhancements to both the test and evaluation sets. From these results we can safely conclude the reliability of the performance to unseen data as the evaluation results are fairly consistent. Context-dependence gives a huge average advantage of around 11%. VTLN, manual corrections of lexical entries and post-filtering of evaluation hypotheses based on frequency errors bring down the WER at a current best performance of 25.3% for the evaluation set. The mappings in the post-filter were decided based on initial substitution error reports. Included in the mapping are the following:

Table 5 Best WERs of Context Independent System from Successive Label Writing (on Dev Set).

Labeling epoch	**LDA only**	**OFS**
1	34.8	31.5
2	**33.8**	**30.1**
3	34.0	30.3
4	34.3	31.0
5	34.4	31.0
6	34.6	30.7
7	35.0	30.9

Table 6 Best WERs of Context Dependent System from Successive Label Writing (on Dev Set).

Labeling epoch	**2000**	**3000**
1	22.5	22.0
2	**22.2**	**21.8**
3	22.3	22.0
4	22.7	22.2
5	23.4	22.5
6	23.7	22.7
7	24.1	23.2

Table 7 Performance (WER) of trained systems.

Description	**Dev**	**Eval**
Context-independent	30.0	35.3
Context-dependent	21.7	27.0
VTLN	21.0	25.9
Manual corrections	**20.5**	**25.3**

1. Words in Tagalog that are closely pronounced to their English counterparts are mapped to English.
2. Most common spelling variants are mapped to a single consistent spelling.
3. Affixated words with simple root extraction (e.g. *pinaka* + [adjective]) are separated from their affixes. Cases of infixes and more complex inflections are not touched.
4. Contractions are expanded.

6. Misrecognition Analysis

6.1 General Observations

Using the system that generated the best performance from Table 7, we analyzed the results of the scoring. On average (development and evaluation sets), the distribution of substitutions, insertions, and deletions are roughly at 70%, 15%, and 15%, respectively. A summary of average relative contributions of different word types for insertion and deletion errors can be found in Table 8. Note that because there are certain words that can be considered as both a Tagalog or an English word, the total when added up exceeds 100%. Error analysis reports currently do not provide the context of the error occurrence. Based on this summary, loan words contribute up to 26.7% of insertion and deletion errors, which is about 7.8% of the total WER.

We then compared the best system with the first FSA-based baseline system in terms of the absolute contribution of loan words to the WER. From Table 9, we can say that data-driven approach solved roughly 44.8% and 6.6% of the contributions of Tagalog and loan words to WER, respectively. For our systems, this is a 3% absolute increase in accuracy for loan words. This is based on a database that contains 20% of loan words in the lexicon and roughly 40% in the actual prompts.

6.2 Error Trends

Aside from the fact that the majority of errors are substitution types, the frequency of occurrence and the acoustic relationship between error pairs give more meaningful insights. Based on the evaluations, we have defined six major trends of substitution errors. The rank and relative contributions of these are summarized in Table 10. From this table, substitution errors that contain loan words on average contribute roughly 7.4% to the total WER. Notable trends in the errors that cannot be simply solved using post-filtering of the hypotheses are mostly morphological in nature. A majority of similar onset errors contain suffixations of *-ng* to nouns that do not contain them (*e.g. ano* becomes *anong*). There is also a prevalence of reduplication of initial and middle syllables for verbs (*e.g. inisip* becomes *iniisip*). Homonyms are mostly concentrated on loan words and can be attributed to poor coverage of the language model. Homonyms can also span more than a single word. Several examples in Tagalog are: *naaakyat* (possible to go up) vs. *na aakyat* (that is going up), *naman niyang* vs. *naman 'yang* (totally different meanings). Orthographical errors due to spelling variants can be alleviated by creating post-filters that can assign a single reference word to multiple entries that have the same connotation. Miscellaneous errors come from acronyms and non-speech sounds.

To further investigate the nature of the errors, we separated the decoding of the test set utterances based on the language characteristics mentioned in Sect. 2. For *code-switching*, we used a list of loan words to flag each utterance. However, utterances with mere proper noun usage are not considered as a switch. Flagging of utterances with inflections is based on the 80 known affixation rules in Filipino. Particles usage is based on the following words: *na*, *pa*, *man*, *nga*, *din*/*rin*, *lang*, *naman*, *daw*/*raw*, *po*/*ha*, *ba*, *pala*, *muna*, *yata* and some cases of *kaya*, *tuloy*, *kasi*, *sana*. For flagging usage of non-standard orthography, we used the results from [25] for the list of spelling variants. However, we did not explicitly include the counts due to the fact that almost all utterances were being included in the set and those being left out from the smaller set are very short phrases. Tables 11 and 12 give us the result of these counts.

We used the optimal setting for the development set to evaluate all the divided sets. Tables 13 and 14 show the achieved WERs with and without post-filtering. It is evident that post-filtering the hypotheses gives an average of 0.62% absolute improvement for all systems. Based on these re-

Table 8 Average relative percentage of word types to insertions and deletions (Pre-filtering).

	Insertions	Deletions
Tagalog	74.6	82.9
Loan words	34.5	17.9
Acronyms	0.6	4.0
Nonspeech	1.1	1.0

Table 9 Average absolute percentage contributions to WER (in parentheses) of Tagalog versus loan words (development set only).

	Baseline (51.5%)	Best System (22.9%)
Tagalog	40.1	17.0
Loan words	10.4	7.0

Table 10 Substitution error types and trends in order of relative frequencies of occurence (Pre-filtering).

Description	Frequency	Example
Tagalog-Tagalog	65.5	*mula* vs. *wala*
Tagalog-Loan	22.0	*atensyon* vs. *retention*
Loan-Loan	14.0	*adjust* vs. *jazz*
Similar ending	14.4	*wrap* vs. *sarap*
Similar onset	12.9	*akin* vs. *aking*
Homonyms	9.6	*bakit* vs. *bucket*
Different middle	4.5	*buto* vs. *boto*
Orthographical	3.6	*kaunti* vs. *konti*
Others	3.2	*T.V.* vs. *T.B.*

Table 11 Ratio of utterances with and without specific language conditions.

Set Condition	Train	Dev	Eval
All utterances	33340	1523	1529
Code-switching	7308 : 26032	552 : 971	560 : 969
Inflections	26886 : 6454	1194 : 329	1205 : 324
Particles usage	17458 : 15882	766 : 757	774 : 755

Table 12 Ratio of number of reference words per set.

Set Condition	Dev	Eval
All utterances	20272	19406
Code-switching	11804 : 9669	10131 : 10460
Inflections	18341 : 2730	17635 : 2763
Particles usage	14180 : 7209	13280 : 7355

Table 13 WERs of specific language condition sets, pre-filtering stage.

Condition	Dev	Eval
All utterances	21.0	25.9
Code-switching	24.2 : **17.7**	29.7 : **22.8**
Inflection	**20.0** : 31.0	**24.9** : 35.0
Particles usage	**20.9** : 22.0	**25.7** : 27.6

Table 14 WERs of specific language condition sets, post-filtering stage.

Condition	Dev	Eval
All utterances	20.5	25.3
Code-switching	23.7 : **17.1**	29.0 : **22.2**
Inflection	**19.1** : 30.4	**24.1** : 34.4
Particles usage	**20.3** : 21.5	**25.0** : 27.1

sults, we have the following observations:

1. Generally speaking, around 20% average absolute contribution to the WERs come from uncorrected human errors in the data both from the recordings and the transcriptions, differing accents, and those enumerated in Table 10.
2. For code-switching, despite utterances flagged as non-switching being higher in number, it achieved a lower WER due to a lesser number of loan words. Loan words contributed about 6.4% and 0.5% average absolute WER to the switching and non-switching sets, respectively.
3. Utterances with inflections had the highest gain from post-filtering due to the large contribution of Tagalog words. Some inflected words still contributed around 1.82% and 0.73% average absolute WER to the respective cases. These generally come from inflected reference words with reduplicated middle syllables substituted with no reduplication (e.g. *makakaramdam* vs. *makaramdam*). Note however that 9% absolute WER for utterances without inflections come from loan words.
4. The 1.5% average relative advantage of utterances without particles and auxiliary words mostly come from loan words and the general errors. The average absolute contributions of the particles to the WERs are at 0.85% and 0.26%, respectively. Note the difference in number of reference words.

Finally, we investigated the quality of the context-dependent models being generated. Due in part to the large number of context-dependent models and the influence of the language model, our analysis was based on the recognition of the training set itself and the language model scores were discounted from the decoding. Not surprisingly, as was observed from the pre-analysis stage of our development, n and ŋ are the most misrecognized. As for vowels, the o and u sounds are the most interchanged. For our systems, solving the distinction problem between n and ŋ can provide around 3.84% absolute increase in recognition accuracy. The vowels u and o accounts for 2% of our systems WER. All other phones are of average contributions to the error. Some however are insignificant in contribution due to underrepresentation. These phones are (in increasing order of representation): z, ɲ, l̪, ð, and θ. This reflects a subset of the loan words that is accountable for almost 6% of our system's WER.

7. Broadcast News Captioning Application

The developed system has been successfully demonstrated in an offline video transcribing system, with the vocabulary and language model adapted specifically for broadcast news. This system makes use of downloaded news videos in Filipino from YouTube and transcribes the videos from a constantly applied segment on the nearest silent point after a predetermined constant duration. The system performs fairly well, except for spontaneous interviews and speeches. Using a small evaluation set of 21 videos from different television networks, the recognition accuracy was found at 78.3%. The system is currently used for transcribing files for manual correction later, facilitating the creation of more training data for closed-captioning applications in Filipino[†].

8. Conclusion

We have described the ongoing development of a continuous Filipino speech recognition system in detail, addressing several practical issues that had an immense effect on previously reported evaluation scores. The creation of a new, richer speech database was first prioritized to address earlier data sparsity issues. Aside from a more careful selection of training and testing sets, text data were further analyzed to lessen any inherent bias. After several training and enhancements, we were able to achieve the lowest word error rate at 20.5%, using a context-dependent system that can still perform in real time. This is at par with the reported accuracy for Filipino speech, notwithstanding the fact that our system allows for code switching that contributes considerable difficulty in the recognition. With an open-domain language

[†]Samples can be seen from videos on the following URL: http://www.youtube.com/user/UPDSPISIP

model, speaker-independent acoustic models, and a more reliable evaluation performance near 80% word recognition accuracy, the data-driven baseline system can be considered as a stepping-stone towards practical application.

The next steps in development can be derived from the challenges presented in the first part of this paper and the impact they made based on the results of the evaluation experiments. For example, better language models can be produced by taking morphology into consideration. Handling of inflections can be refined by establishing affixation rules that can and cannot be broken down into separate words for both the language model and the hypothesis post-filter. Some phone models also need special attention, especially for distinguishing between words ending in n or ŋ, which accounts for 3.84% absolute WER for our systems. The post-filtering techniques for spelling variants can also be further improved by including more homonyms that doesn't largely affect the context, especially for later purposes such as for machine translation. Also for our systems, proper handling of loan words can have absolute accuracy gains of up to 6%. Since in practice most decisions made in development is based on empirical findings, more experiments must be done to completely map the evaluation trajectory. It is expected that from the provided set of challenges, other researchers can contribute their own set of solutions for a richer development community working on Filipino ASR.

Acknowledgments

The authors would like to thank the Philippine Council for Industry, Energy and Emerging Technology Research and Development of the Department of Science and Technology (PCIEERD-DOST) and the Digital Signal Processing Laboratory of the University of the Philippines, Diliman for the ISIP Tagalog Speech Database.

References

[1] F. Ang, J. Ancheta, K. Francia, and K. Chua, "Evaluation of smoothing techniques for language modeling in automatic Filipino speech recognition," TENCON 2012 - 2012 IEEE Region 10 Conference, pp.1–5, 2012.

[2] F. Ang, M. Burgos, and M. De Lara, "Automatic speech recognition for closed-captioning of Filipino news broadcasts," 2011 7th International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp.328–333, 2011.

[3] R. Sagum, R. Ensomo, E. Tan, and R. Guevara, "Phoneme alignment of Filipino speech corpus," TENCON 2003, Conference on Convergent Technologies for the Asia-Pacific Region, vol.3, pp.964–968, 2003.

[4] S. Sakti, R. Isotani, H. Kawai, and S. Nakamura, "The use of Indonesian speech corpora for developing Filipino continuous speech recognition system," Proc. O-COCOSDA, pp.56–61, Nov. 2010.

[5] R. Nolasco, "Filipino and Tagalog, not so simple," http://www.dalityapi.com/2007/08/articles-filipino-and-tagalog-not-so.html, 2007. Accessed June 2013.

[6] M. Tayao, "The evolving study of Philippine English phonology," World Englishes, vol.23, no.1, pp.77–90, 2004.

[7] J. Wolff, "Tagalog," in Encyclopedia of language & linguistics, Elsevier, New York, 2006.

[8] M. Bautista, "Tagalog-English code switching as a mode of discourse," Asia Pacific Education Review, vol.5, no.2, pp.226–233, 2004.

[9] "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict. Accessed: June 2013.

[10] K. Rara, F. Cristobal, F. De Leon, G. Zafra, C. Clarin, and R. Guevara, "Towards the standardization of the Filipino language: Focus on the vowels of English loan words," International Symposium on Multimedia and Communication Technology (ISMAC), 2009.

[11] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997, ICASSP-97, pp.1039–1042, vol.2, 1997.

[12] T. Kaukoranta, P. Fränti, and O. Nevalainen, "Iterative split-and-merge algorithm for VQ codebook generation," Optical Engineering, pp.2726–2732, 1998.

[13] M.J.F. Gales, "Semi-tied covariance matrices for hidden Markov models," IEEE Trans. Speech Audio Process., vol.7, no.3, pp.272–281, 1999.

[14] H. Yu and A. Waibel, "Streamlining the front end of a speech recognizer," ICSLP 2000, pp.353–356, 2000.

[15] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at sixteen: Update and outlook," Proc. IEEE Automatic Speech Recognition and Understanding Workshop, Dec. 2011.

[16] E. Ristad, "A natural law of succession," tech. rep., Comp. Sci. Dept., Princeton Univ., CS-TR-495-95, 1995.

[17] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recognizer," IEEE Trans. Acoust., Speech Signal Process., vol.35, no.3, pp.400–401, 1987.

[18] I.H. Witten and T. Bell, "The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression," IEEE Trans. Inf. Theory, vol.37, no.4, pp.1085–1094, 1991.

[19] H. Ney and U. Essen, "On smoothing techniques for bigram-based natural language modelling," 1991 International Conference on Acoustics, Speech, and Signal Processing, 1991, ICASSP-91, vol.2, pp.825–828, 1991.

[20] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," 1995 International Conference on Acoustics, Speech, and Signal Processing, 1995, ICASSP-95, vol.1, pp.181–184, 1995.

[21] S.F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," Proc. 34th Annual Meeting on Association for Computational Linguistics, ACL '96, pp.310–318, Association for Computational Linguistics, 1996.

[22] M. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Computer Speech and Language, vol.12, pp.75–98, 1998.

[23] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," Proc. ASRU, 2001.

[24] "NIST multimodal information group speech tools." http://www.nist.gov/speech/tools. Accessed: June 2013.

[25] J. Ilao and R.C. Guevara, "Investigating spelling variants and conventionalization rates in the Philippine national language's system of orthography using a Philippine historical text corpus," in Proc. of O-COCOSDA, Dec. 2012.

Federico Ang was born in Manila, Philippines, on May 14, 1986. He received his B.S. and M.S. degrees from the University of the Philippines, Diliman in 2007 and 2009, respectively. He is currently a doctor student at the Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Japan. His research interests are in the areas of speech signal processing and recognition, and machine learning for signal processing systems. He is an IEEE member since 2010.

Rowena Cristina Guevara is a professor at the Electrical and Electronics Engineering Institute, University of the Philippines (Diliman) and is affiliated with the UP Digital Signal Processing Laboratory. She finished her Ph.D. at the University of Michigan, Ann Arbor in 1997. Her areas of specialization include speech and audio signal processing, time-frequency analysis and synthesis, and artificial intelligence. She was recently appointed as the new executive director of the Philippine Council for Industry, Energy and Emerging Technology Research and Development (PCIEERD) of the Philippines' Department of Science and Technology (DOST).

Yoshikazu Miyanaga was born in Sapporo, Japan, on December 20, 1956. He received his B.S., M.S., and Dr. Eng. degrees from Hokkaido University, Sapporo, Japan, in 1979, 1981, and 1986, respectively. He is currently a Professor at the Graduate School of Information Science and Technology, Division of Media and Network Technologies in Hokkaido University. His research interests are in the areas of speech signal processing, LSI design with low power consumption, and green systems of wireless communication. Dr. Miyanaga is an IEICE fellow. He is the vice-president of Asia-Pacific Signal and Information Processing Association (APSIPA) from 2009 to 2013. He was a distinguished lecturer (DL) of the IEEE CAS Society (2010–2011) and he is now a Board of Governor (BoG) of the IEEE CAS Society (2011–2013).

Rhandley Cajote is an associate professor at the Electrical and Electronics Engineering Institute, University of the Philippines (UP Diliman). He finished his B.S. and M.S. in Electrical Engineering at UP Diliman. He received his Ph.D. in Electrical Engineering from Chulalongkorn University, Department of Electrical Engineering. His research interests are in the area of signal and image processing, video coding, video communications, and digital signal processors. He is a member of IEICE and IEEE.

Joel Ilao is an assistant professor in the Computer Technology Department, College of Computer Studies at De La Salle University. He finished his Ph.D. at the University of the Philippines, Diliman, working on a corpus-based analysis of the Filipino written text. His research interests include computational linguistics, digital signal processing, and machine vision.

Michael Gringo Angelo Bayona is an instructor at the Electrical and Electronics Engineering Institute of the University of the Philippines (UP Diliman), and is affiliated with the UP Digital Signal Processing Laboratory. He finished his B.S. in Electronics and Communications Engineering and M.S. in Electrical Engineering at UP Diliman, developing a reading tutor for young learners of the Filipino language. His research interests include speech and audio signal processing and machine learning.

Ann Franchesca Laguna received her B.S. in Computer Engineering in 2012 from the Electrical and Electronics Engineering Institute of the University of the Philippines, Diliman where she has been a student affiliate of the Digital Signal Processing Laboratory since 2009. Her research interests are in mobile applications programming and in audio and speech signal processing. She is currently an M.S. Electrical Engineering Student in the same university.