Philippine Journal of Science
144 (1): 81-89, June 2015
ISSN 0031 - 7683

Development, Implementation and Testing of Language Identification System for Seven Philippine Languages

Ann Franchesca B. Laguna[1] and **Rowena Cristina L. Guevara**[2]

[1]Computer Technology Department, College of Computer Studies,
De La Salle University, Taft Avenue, Manila
[2]Electrical and Electronics Engineering Institute,
University of the Philippines Diliman, Quezon City

Three Language Identification (LID)approaches, namely, acoustic, phonotactic, and prosodic approaches are explored for Philippine Languages. Gaussian Mixture Models (GMM) is used for acoustic and prosodic approaches. The acoustic features used were Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), Shifted Delta Cepstra (SDC) and Linear Prediction Cepstral Coefficients (LPCC). Pitch, rhythm, and energy are used as prosodic features. A Phone Recognition followed by Language Modelling (PRLM) and Parallel Phone Recognition followed by Language Modelling (PPRLM) are used for the phonotactic approach. After establishing that acoustic approach using a 32nd order PLP GMM-EM achieved the best performanceamong the combinations of approach and feature, three LID systems were built: 7-language LID,pair-wise LID and hierarchical LID; with average accuracy of 48.07%, 72.64% and 53.99%, respectively. Among the pair-wise LID systems the highest accuracy is 92.23% for Tagalog and Hiligaynon and the lowest accuracy is 52.21% for Bicolano and Tausug. In the hierarchical LID system, the accuracy for Tagalog, Cebuano, Bicolano, and Hiligaynon reached 80.56%, 80.26%, 78.26%, and 60.87% respectively. The LID systems that were designed, implemented and tested, are best suited for language verification or for language identification systems with small number of target languages that are closely related such as Philippine languages.

Key Words: Gaussian Mixture Models, Language Identification, Speech Processing

INTRODUCTION

Language Identification (LID) systems (Ambikairajah 2011; Schultz 2006; Benesty 2007) recognize the language of a speech signal of unknown language.The identified language is chosen based on the similarity of the speech signal with statistical models of the language. LID systems are used for multilingual speech processing systems such as automatic speech recognition systems (ASR), speech to speech translators, reading tutors, and automatic call routing. Most language identification research focus on multinational languages that are less perceptually and lexically similar than Philippine languages. An LID system for seven Philippine languages is designed in this research. The languages included in the LID system are Tagalog (TGL), Cebuano (CEB), Hiligaynon (HIL), Kapampangan (KAP), Bicolano (BCL), Waray (WAR), and Tausug (TSG).

The most intuitive approach in determining the language is by using lexical and grammatical rules. Certain words and word sequences appear in one language and not in others. This technique, though simple for humans, however require time-consuming data preparation and labor-intensive processing on computers when considering speech data. Another

*Corresponding author: ann.laguna@dlsu.edu.ph

approach is to determine the speaker's language based on perceptual properties of a speech signal. Knowledge of the vocabulary and syntax is therefore no longer a prerequisite. This approach in which the perceptual quality of speech is utilized, instead of basing the language decision on the words, is used in LID systems. The latter approach is used for this research.The performance of LID systems is measured in terms of correct recognition of a language and is usually expressed in terms of recognition rate or accuracy.

Data Collection and Preparation

The Digital Signal Processing (DSP) Laboratory of the University of the Philippines Diliman developed anadult Philippine language speech database.The speech database consists of prompted words and utterances with fixed vocabulary, as well as spontaneous speech. The data is recorded at 16 kHz sampling rate and 16 bit quantization and then resampled at 8kHz. There are two recording setups: the laboratory setup and the field setup.The recordings for the laboratory setup is done in a pseudoanechoic chamber using a multipattern condenser microphonevia a multi-channel analog mixer and *preamp* with a built-in audio interfaceconnected to a desktop computer. The field setup uses a *headset* with noise-cancelling microphone connected to a laptop and is usually done in quiet rooms in different provincial locations. Most of the Tagalog and Tausug as well as a significant amount of Cebuano and Hiligaynon are recorded using the laboratory setup. The rest of the languages are mostly recorded using the second setup. There are 200 speakers for each language: 100 male and 100 female.The experiments implemented in this study use the spontaneous speech in the corpora. Table 1 shows the number of hours for the train set and test set of the spontaneous data. Each spontaneous speech file is 10-40 seconds long.

Table 1. Number of Hours of Spontaneous Data.

	Train Set (hours)	Test Set (hours)
TGL	4.3	1.9
CEB	4.1	1.8
BCL	4.8	2
KAP	4.3	2.6
HIL	2.8	1
WAR	4.7	2.8
TSG	2.4	1

Acoustic and Prosodic Language Identification using Gaussian Mixture Models

Acoustic and prosodic features are widely used for speech processing systems. Acoustic features, such as Mel Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), Shifted Delta Cepstra (SDC) and Linear Prediction Cepstral Coefficients (LPCC) are based on the spectral characteristics of the speech signals. Prosodic features such as tone, stress and rhythmare usually modelled using pitch or the fundamental frequency f_0, energy, and the normalized duration of phonemes and syllables. These acoustic and prosodic features are commonly used for LID systems. (Wong 2004)

A segmentation approach was used in extracting the front-end acoustic features. Afterwards, an energy based Voice Activity Detection (VAD) was used to remove silent frames. Feature vectors were then extracted from each frame. The acoustic features of the existing speech corpora were extracted for each frame with window length of 25 milliseconds and frame shift of 10 milliseconds.Twenty filter bank channels are used to compute for the MFCC and PLP. The speech segment is band limited to 300-3,700 Hz. Cepstral Mean Normalization (CMN) is done per utterance. The mean of each utterance is computed and subtracted from each frame of that utterance. The number of coefficients for the different feature vectors is13 for MFCC, 9 for PLP, 32 for SDC, and 9 for LPCC.

The computation of SDC has four main parameters: M, D, P and K. M specifies the number of MFCC values to use. D and P represent the difference value and the number of frames for each delta calculation, respectively. K is the total number of SDC values.The SDC configuration of 7-1-3-7 (Torres-Carrasquillo 2003) is used for this research, which uses 21 frames to obtain temporal data for 210 ms while retaining the fine-grained information within those frames.

Gaussian Mixture Models

Acoustic and prosodic language identification uses Gaussian Mixture Model (GMM) to create a statistical model of a language (Ambikairajah 2011; Schultz 2006; Benesty 2007). GMM represents the language as a linear combination of multivariate Gaussian distributions. Each of the Gaussian distribution is represented by its mean and variances as shown in Equation 1.

$$g(x) = \sum_{i=1}^{N} \lambda_i N(x,; \mu_i; \Sigma_i) \qquad (1)$$

Where g(x) is the GMM, N is the number of mixture components, x is the feature vector, λ_i is the weight of the ith Gaussian mixture, m_i is the mean vector of the ith Gaussian mixture and Σ_i is the covariance matrix of the ith Gaussian mixture

The Expectation Maximization (EM) Algorithm (Ambikairajah 2011) estimates each language as a separate GMM. The signal is then compared to each of the language GMM. Another approach, the Universal Background

Model (UBM)(Ambikairajah 2011) approach, uses a single GMM, the background model, to represent the feature vectors from all the languages, in addition to the EM method. The GMM is then adapted using Bayesian Adaptation for each of the language. To obtain the language of the unknown speech utterance, the posterior probabilities of a series of speech feature vector is obtained for each language GMM. The GMM with the largest maximum posterior probability is chosen.

Experiment on including Read Speech for the LID Training and Test Set

Read speech have different acoustic and prosodic characteristics compared to spontaneous speech. This experiment investigates the appropriateness of including read speech in the LID database. Two 2nd order GMM language sets are created; one set with spontaneous speech, and the other set with read and spontaneous speech. Both sets use Tagalog (TGL), Cebuano (CEB), and Bicolano (BCL) language. Table 2 shows the recognition rates for this experiment. The highest accuracy is obtained using PLP of spontaneous speech only, with an accuracy of 76.77%. The inclusion of read speech in the training lowers the language recognition, despite having more data samples, as compared to using only spontaneous speech. The lower recognition rate shows that the languages are highly similar in read speech as compared to spontaneous speech only. Because of this similarity the complexity of recognizing the languages increases. The results of this experiment affect the succeeding experiments where only the spontaneous speech part of the corpora is used in creating the LID system.

Table 2. Recognition Rate of Read and Spontaneous Speech for LID systems for different features

	Read and Spontaneous	Spontaneous Only
MFCC	43.87%	45.70%
PLP	37.56%	76.77%
SDC	40.29%	49.34%
LPCC	27.32%	62.89%

Seven-Language GMM-EM and GMM-UBM LID System

GMM-EM and GMM-UBMare compared with each other by creating a seven-language LID system. Severalnumber of Gaussian mixtures were tested to determine the number of Gaussian mixtures that model the language more accurately. Higher number of mixtures models the training set more accurately, this however may cause overfitting of the training data. Overfitting occurs when the training set accuracy increases but the test set accuracy no longer improves. Higher ordered GMMs also need longer training time than the lower-ordered GMMs. Hardware limitations dictatethe speed of training a GMM. This limitation is based on the length of the feature vectors or the dimension of the GMM, the number of frames or size of data for each language or the number of training samples, and the number of mixture components or the order of the GMM.

From Table 3, the best performance is obtained for the LID system using the PLP feature followed by MFCC. LPCC and SDC do not perform comparatively well for the LID task. This shows that PLP and MFCC contain more discriminative information for Philippine languages as compared to LPCC and SDC.With SDC having 32 parameters, the GMM is having a hard time to converge to a solution because the available data becomes sparser. Increasing the amount of data or changing the SDC parameters can improve this situation.The GMM-UBM approach achieves a comparable accuracy compared to the standard GMM-EM approach. The GMM-UBM approach provides a faster training time because the Universal Background Model provides initial conditions for the GMMs that arecloser to the speech characteristics of the languages. Because of memory limitations, only 20% of the training set is used to model the UBM. Increasing this would improve the UBM but increase the memory requirements and training time.

Table 3. Comparison of Accuracy (in percent) for GMM-EM and GMM-UBM Algorithm for the different acoustic feature vectors

# of Mixtures	MFCC		PLP		LPCC		SDC	
	EM	UBM	EM	UBM	EM	UBM	EM	UBM
1	23.15	14.29	18.70	14.29	14.29	14.29	14.29	14.29
2	31.47	28.36	37.74	40.70	14.75	13.66	14.29	14.29
4	25.80	29.51	46.31	40.70	16.44	15.20	14.29	14.29
8	31.59	29.09	47.46	47.04	18.95	21.13	15.94	15.15
16	35.86	34.32	48.08	43.32	20.58	18.63	18.68	16.37

The highest accuracy for this experiment is obtained using 16 mixture PLP GMM-EM LID systems. Table 4 shows the confusion matrix of predicted and actual language in this system. The average accuracy of this system, taken from the diagonal of the confusion matrix, is 48.07%. Tagalog, Cebuano, and Hiligaynon have the highest recognition rates among all the languages. Bicolano and Tausug also provide good recognition rate however Bicolano is commonly misclassified as Kapampangan and Tausug is misclassified as Waray. Kapampangan is commonly misclassified with either Bicolano or Cebuano. Waray on the other hand performed poorly for this LID system. Thus, this system is good for recognizing Tagalog, Cebuano and Hiligaynon. If Bicolano and Kapampangan are treated as a single language group the recognition for

Table 4. Confusion Matrix for 16 mixturePLP GMM-EM for 7 Languages

		Actual Language						
		TGL	**CEB**	**BCL**	**KAP**	**HIL**	**WAR**	**TSG**
Predicted Language	TGL	82.24	21.53	11.11	2.41	3.26	0.00	0.00
	CEB	3.95	70.14	0.56	18.07	9.78	1.21	1.09
	BCL	0.66	2.08	42.78	27.11	3.26	21.21	3.26
	KAP	1.32	0.00	30.56	33.73	7.61	47.88	6.52
	HIL	1.97	0.00	2.78	9.04	67.39	14.55	13.04
	WAR	0.00	0.69	2.22	4.22	0.00	5.45	41.30
	TSG	9.87	5.56	10.00	5.42	8.70	9.70	34.78

Bicolano-Kapampangan increases. Most of the predicted Waray, are Tausug prompts. Hence, using the Waray GMM, for Tausug would also increase the recognition for Tausug. This LID system cannot be used effectively in detecting Waray.

Figure 1 shows the effect of varying the number of Gaussian mixtures on the recognition rate of the train and test set for the PLP GMM-UBM LID system. The recognition rate for the training set continually increases as the number of mixtures increases. Increasing the number of mixtures beyond 32 does not provide anysignificant amount of improvement in the accuracy and would only increase the computational complexity of the system. Hence, the optimal number of mixtures is 32.

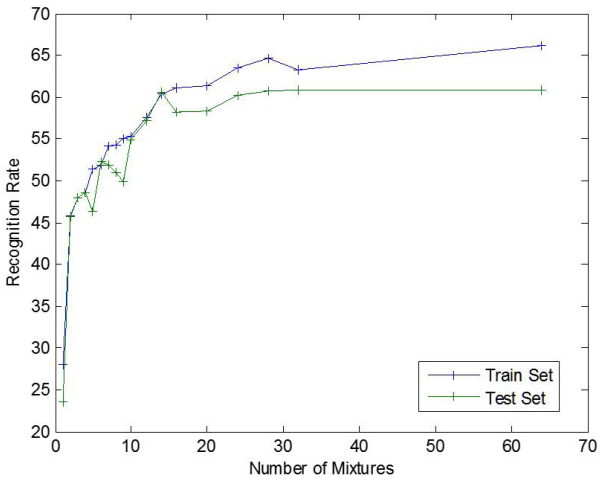

Figure 1. Effect of varying the number of Gaussian Mixtures for the PLP GMM-EM LID System

Two-Language GMM-EM LID System

Since the number of languages affects the complexity and performance of the system, a pair-wise evaluation is explored. Two hypothesis languages and a speech utterance are taken, then a decision is made on which of the two languages the speech segment belongs to. The acoustic features are extracted from the speech utterance. This is then taken as an input to the two chosen language GMMs.

Table 5 shows the performance of the LID systems created for the two language tests. Tagalog – Hiligaynon pair has the highest accuracy at 92.23% and Bicolano – Tausug has the lowest recognition rate at 52.21%. The average accuracy for the 21 pair-wise LID systems is 72.64%. Those languages paired with Tagalog and Cebuano have higher accuracy compared to other languages. The created LID systems are thus good for discriminating between Tagalog, Cebuano and other languages. The small exception would be discriminating between Tausug and Tagalog which only gives an accuracy of64.60%. Generally, a pair-wise system has higher accuracy than that of the 7–language LID system with only 48.08% accuracy. This shows that recognition for smaller number of languages provides better accuracy than when the discrimination is done for seven languages.

Table 5.Two Language Test LID System Test Set Recognition Rate.

1st Language	2nd Language	Accuracy
TGL	CEB	82.24%
TGL	BCL	83.78%
TGL	KAP	88.38%
TGL	HIL	92.23%
TGL	WAR	86.37%
TGL	TSG	64.60%
CEB	BCL	82.85%
CEB	KAP	79.94%
CEB	HIL	86.99%
CEB	WAR	90.20%
CEB	TSG	80.83%
BCL	KAP	66.70%
BCL	HIL	69.36%
BCL	WAR	58.79%
BCL	TSG	52.21%
KAP	HIL	59.20%
KAP	WAR	55.35%
KAP	TSG	57.26%
HIL	WAR	56.11%
HIL	TSG	67.39%
WAR	TSG	64.70%

Seven-Language GMM-EM using Prosodic LID System

Prosody refers to the patterns of intonation, rhythm and accent in speech. The most basic prosodic features are measured in terms of energy and pitch contours as well as rhythm and duration. A meaningful pattern of pitches or intonation patterns are considered crucial in Philippine prosody. The intonation pattern is described using three features: pitch points, pitch levels and pitch contour.The features are first extracted from the speech data using a phonetic speech analyzer (Boersma 2014). The features are then passed to a GMM classifier for training the sequence. The language model is then used to classify a testing set disjoint from the training set.

Table 6 shows the results of using Gaussian Mixture Models. The performance of using prosodic features performed poorly at 17.22% accuracy as compared with using acoustic features which has an accuracy of 48.08%. However,prosodic LID systems have high accuracy of 99.75% when the training set is used. As the number of mixtures increases, training set accuracy increases but the test set accuracy does not improve. This is an overfitting problem in which the created model is able to fully capture the necessary features in the training setbut not for unseen data. Hence, the created model does not fully characterize the language. One of the assumptions in extracting the prosodic features is that each utterance is one prosodic unit. This however does not hold for all the utterances. Prosodic transcription should be done to mark the boundaries of the prosodic units in the system.

Table 6. Prosodic Language Identification train and test set accuracy (in percent) for 5 and 10 Legendre polynomials

GMM Order	Legendre Polynomial 5		Legendre Polynomial 10	
	Training Set	Test Set	Training Set	Test Set
1	21.65	16.63	24.38	16.98
2	27.07	16.56	27.49	20.75
4	33.44	18.48	40.79	21.12
8	46.29	20.57	57.97	16.55
16	62.29	16.51	75.35	18.50
32	85.58	14.71	91.83	16.73
64	97.28	17.05	89.90	16.98
128	99.75	17.22	86.92	11.11

Language Identification Using Phonotactic Information

Phonemes are the most basic unit of sound in human speech. The probability of a speech signal to be a specific language given the absence and the frequency of occurrence of phonemes and phoneme sequences can help us identify the language (Schultz, 2006).A log-likelihood score of the N-gram language model is computed for each language to determine the probability of the utterance to be a given language. This probability is shown by the equation below:

$$P(\Psi|L_i) = \sum \log Pi\,(\varphi|\ \varphi_{i-1}, \varphi_{i-2}, \varphi_{i-3}, \ldots, \varphi_{i-N}, L_i) \qquad (2)$$

where $P(\Psi|L_i)$ is the probability that a given phoneme sequence ψ occurs given the language model L_i, and φ_{i-1}, $\varphi_{i-2}, \varphi_{i-3}, \ldots, \varphi_{i-N}$ are the phonemes preceding φ.The syllable structure of Philippine languages is commonly of the form consonant-vowel (CV) or consonant-vowel-consonant (CVC) hence a trigram or a bigram model is sufficient.

To design these systems a phoneme recognizer and a language model for each language is needed. The phoneme recognizers are trained using automatically transcribed read and spontaneous speech, while the language model is trained using spontaneous speech only.

Phoneme Recognition

Creating a phoneme recognizer using Hidden Markov Model Toolkit (HTK) (Young 2006)involves several steps as shown in Figure 2. HTK is a toolkit used to build and manipulate HMMs and is commonly used in speech recognition.

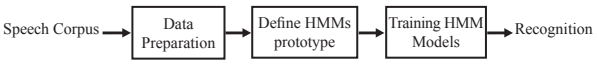

Figure 2. Phoneme Recognizer Design Process.

The data preparation includes collecting speech database, generating transcriptions and dictionaries and defining the phoneme table. A tri-state left-to-right HMM topology is used to represent each phoneme in the acoustic model of the speech recognizer. Re-estimation of the transition probabilities, mean and variance is done using Baum-Welch algorithm. The measures are then re-estimated until convergence. Recognition is implemented using the Viterbi Algorithm.

Mel Frequency Cepstral Coefficients (MFCC) are used to train the phoneme recognizer. Twelve MFCC Parameters and the cepstral mean, appended with the delta and acceleration coefficients are extracted for each speech frame giving a total of 36 coefficients. The window length of each frame is 25 milliseconds with a frame shift of 10 milliseconds.

Each particular phoneme has a corresponding symbol in the phoneme set. For example, the symbol AO corresponds to the vowel “ao” and the symbol “HH” corresponds to the consonant “h”. The phonemes used in this set are either a monophone or a diphthong. A monophone is a

single acoustic unit, while diphthongs such as AW an SY are combinations of two monophones.Using the phoneme set, a pronunciation dictionary is created. Some of the pronunciation in the dictionary is generated automatically with the rule "read as written." For example the word AABOT has the pronunciation A Q A B AO T. The glottal stop /Q/ is inserted because of phonetic rulesthat a syllable must follow a CV or CVC pattern within a word.

The phoneme recognizer relies on the transcription of the speech corpus to determine the boundaries of the phonemes. This is necessary in creating a model for the individual phonemes. The exactness of the transcription therefore is crucial in speech recognition. The phonemic transcriptions used for this research are obtained from the pronunciation dictionary and word/sentence level transcription of the speech data using forced alignment. Forced alignment uses the transcription to determine the time segment in which the pronunciation occurred. Since the phonemic transcriptions are automatically generated, the exactness of the phonetic boundaries are not as exact as hand-labelled transcriptions. Improving the dictionary and hand-labelling the transcriptions are both tedious and time consuming. The availability of skilled transcribers familiar withthe languages is also a challenge.

Table 7 shows the resulting accuracy of the individual phoneme recognizers. The recognition rate averages at 45%. Tagalog and Bicolano have the highest recognition rate at 58.13% and 52.1% while Kapampangan has the lowest recognition rate at 34.63%. The pronunciation in the Tagalog and Bicolano dictionaries are manually checked and hence give more accuracy than the other languages. The low accuracy of the Kapampangan language may be due to the ambient noise present in the recording data and also the manner of speaking of Kapampangan where phonemic sounds can vary from word to word depending on the speaker and the context of the word.

Table 7. Accuracy of Individual Phoneme Recognizer.

Language	Accuracy
TGL	52.1%
CEB	42.86%
BCL	58.13%
KAP	34.63%
HIL	43.99%
WAR	45.5%
TSG	44.45%

Phonotactic LID system using PRLM and PPRLM

Two popular approaches of phonotactic LID: Phone Recognition Followed by Language Modelling or PRLM and Parallel Phone Recognition Followed by Language Modelling or PPRLM (Ambikairajah 2011; Shen 2006), are used in this research. For the PRLM approach, a phoneme recognizer is trained for a specificlanguage. The phoneme recognizer is trained using both read and spontaneous speech. The phoneme and phoneme sequences obtained from the phoneme recognizers are then used to create phonotactic language models. A phonotactic language model is trained for each of the languages using only the spontaneous speech in the corpora. For the PPRLM approach, a separate phoneme recognizer is created for each of the languages. The PPRLM phonotactic language models are created similar to PRLM.

By using a single phoneme recognizer for PRLM, the phonemes of the individual languages are assumed to be similar with each other. The obtained phoneme sequences of spontaneous speech of the languages are used to train the phonotactic language models. Table 8 shows the summary of the accuracy of the LID system using different phoneme recognizers. An average of 39% can be observed from the PRLM LID system. Though the phoneme recognition of Bicolano and Tagalog is higher than that of the other languages, the accuracy of the LID task is still comparable to the other languages. Kapampangan, however, which has the lowest phoneme recognition rate, also has the lowest PRLM training set. Phoneme recognition rate therefore is only one of the factors that affect PRLM accuracy. Since a single phoneme recognizer is being used, even if the input speech is a different language, i.e. using a Tagalog recognizer even if the input speech is Tausug, the nuances of the phonemes of the input speech would not be characterized more accurately.

Table 8. Accuracy (in percent) of PRLM using different Phoneme Recognizers.

Phoneme Recognizer	Phoneme Recognition Rate	PRLM Accuracy (Training Set)	PRLM Accuracy (Testing Set)
TGL	52.1	39.17	36.64
CEB	42.86	43.25	44.41
BCL	58.13	40.68	35.70
KAP	34.63	31.24	32.12
HIL	43.99	41.60	43.71
WAR	45.5	39.01	36.67
TSG	44.45	42.44	43.76

The obtained recognition rate of the PPRLM network is 36.62%. Compared to using a single phoneme recognizer with an average of 39% accuracy; there is no significant improvement in the accuracy of the PPRLM framework.

Hierarchical Language Identification

Grouping the most similar languages together can reduce the complexity of the system and increase the accuracy of the LID system. Research using statistical signal processing methods (Ambikairajah 2011), on the actual grouping still need to be explored.In a hierarchical LID framework,a tree network is used to group similar languages together. Only discriminative features are considered in each level hence a different set of features and classifiers can be used for each level. Distance measurements obtained from an unsupervised clustering method is used to determine the language grouping. Distance at higher levels of the hierarchy, are larger than the distance between the language groups in the lower levels. During the identification phase, the language is determined by passing through different LID systems in our tree network.

Figure 3 shows the hierarchical LID classification of the Tagalog test set. In the second tier, 86.84% of the data were classified as Tagalog or Cebuano, while the rest of the data were classified as one of the five other languages. Then on the third tier, 80.26% were classified as Tagalog. From these figures, the hierarchical model generally increases the accuracy of certain languages such as Tagalog, Cebuano and Tausug which are languages on the third tier, while the languages from the lower tiers have lower recognition rates such as Bicolano, Kapampangan and Waray.

The recognition rates for the individual languages are summarized in Table 9. Tagalog, Cebuano, and Tausug good accuracies of 80.26%, 80.55%, and 78.27% respectively while Hiligaynon performs fairly well at 60.87%. Bicolano, Kapampangan, and Waray have low accuracies at 32.78%, 43.98%, and 1.21% using this hierarchical classification. The average accuracy for this

Table 9. Accuracy of Hierarchical Method.

Language	Accuracy
TGL	80.26%
CEB	80.56%
BCL	32.78%
KAP	43.98%
HIL	60.87%
WAR	1.21%
TSG	78.26%

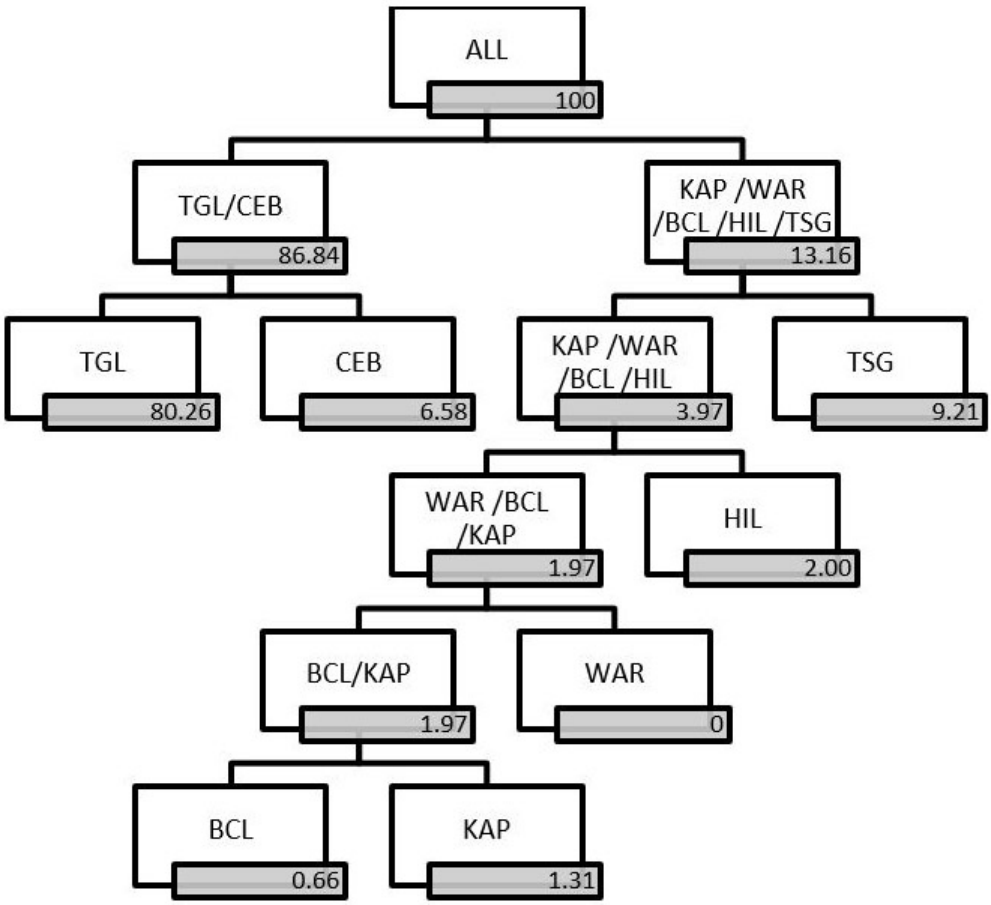

Figure 3. Hierarchical Results using Tagalog as the input speech to the Language Identification system.

hierarchical language identification system is 53.99%. If Waray is not included in the system, the average accuracy is 62.79%. Tagalog, Cebuano, Hiligaynon, and Tausug corpora were recorded mostly in a pseudo-anechoic chamber in the DSP Laboratory while the other three languages were mostly recorded in relatively quiet rooms in the provinces. The recording environment causes the former setup to obtain a much cleaner data with low noise and wider dynamic range.

CONCLUSION AND RECOMMENDATION

The application of acoustic, phonotactic, and prosodic approaches for Philippine Languages is explored in this research. Among the four feature vectors studied, namely MFCC, PLP, SDC, and LPCC, PLP showed the best results. No significant improvement is observed at higher orders and the computation time dramatically increases at GMM orders higher than 32. For the 7-language acoustic LID system the highest average accuracy of 48.08% is obtained using 32nd order Perceptual Linear Prediction GMM.

The prosodic approach using Gaussian Mixture Models and Legendre Polynomials showed high recognition rates for the training set at 99.75% but low recognition rates for the test set at 17.22%. This is a manifestation of overfitting the model to the training data. Increasing the number of prosodic phrases in the data set can further exacerbatethis problem.

The phonotactic approach obtained an accuracy of 43.76% for the PPRLM approach. Automatic transcriptions of the phonemes are used for the training set of the phoneme recognizers. Manual checking of the phoneme transcription by language experts should be done to further improve the accuracy of the phoneme recognizers and thus also improve the phonotactic LID system.

In the hierarchical LID, the average recognition rate is 53.99% and the recognition rates for Cebuano, Tagalog, Tausug and Hiligaynon arethe best among the different languages at 80.55%, 80.26%, 78.26%, and 60.87%, respectively. These four languages were recorded mostly in a pseudoanechoic chamber as compared to Kapampangan, Bicolano and Waray which were mostly recorded in the field. The noise in the field recordings is a possible reason for the low accuracy for the four languages.This problem can be solved in two ways depending on the application. First, only the top four languages are chosen for our LID system and ensure that the input languages are indeed one of the top four languages. Second, if the other languages must be included in the system, a more robust algorithm or approach may need to be investigated further to create a more efficient LID system.

From the experiments, using a pair-wise evaluation approach gives better accuracy than that of using all seven languages. The average accuracy for the 21 pair-wise LID systems is 72.64%. The highest pair-wise accuracy is obtained for Tagalog and Hiligaynon at 92.23%. LID pairs with one language being Cebuano or Tagalog showed better LID recognition rates.

Having prior hypothesis of the actual language of the speech utterance can be used to further improve the performance of the system. The LID system created for this research is best suited for language verification. The pair-wise LID system has the best average accuracy and the pair-wise LID for Tagalog and Cebuanoexhibitedthe best language recognition rates in all the experiments.

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 Laboratoryof the University of the Philippines, Diliman for the ISIP Tagalog Speech Database. This research is also financially assisted by Engineering Research and Development for Technology of DOST (ERDT-DOST).

REFERENCES

AMBIKAIRAJAH E, HAIZHOU LI, LIANG WANG, BO YIN, SETHU V. 2011. Language identification: a tutorial. Circuits and Systems Magazine, IEEE, 11(2):82 –108.

BENESTY J, SONDHI MM, HUANG Y. 2007. Springer Handbook of Speech Processing. NJ, USA: Springer-Verlag New York, Inc., Secaucus.

BOERSMA P, WEENINK D. 2014. Praat: doing phonetics by computer. Retrieved from www.praat.org, on 20 February 2014.

SCHULTZ T, KIRCHHOFF K. 2006. Multilingual Speech Processing. New York: Academic.

SHEN W, CAMPBELL W, GLEASON T, REYNOLDS D, SINGER E. 2006. Experiments with Lattice-based PPRLM Language Identification. In: Speaker and Language Recognition Workshop; 2006. IEEE Odyssey 2006.

SINGER PA, TORRES-CARRASQUILLO TP, GLEASON WM, CAMPBEL L, REYNOLDS DA. 2003. “Acoustic, Phonetic, and discriminative

approaches to automatic language identification. In: Proceedings of Eurospeech; 2003; Geneva, Switzerland, EUROSPEECH p. 1345–1348.

WONG E. 2004. Automatic spoken language identification utilizing acoustic and phonetic speech information," [Ph.D. Dissertation]. June 2014. Speech and Audio Research Laboratory, Queensland University Technology.

YOUNG SJ, EVERMANN G, GALES MJF. 2006. "The HTK Book, version 3.4".