Computer and Information Science
www.ccsenet.org/journal.html

Vol. 1, No. 4
November 2008

Acoustic Pronunciation Variations Modeling for Standard Malay Speech Recognition

Noraini Seman
Faculty of Information Technology and Quantitative Sciences
Universiti Teknologi MARA
40450 Shah Alam, Selangor, MALAYSIA
Tel: 603-5521-1127 E-mail: aini@tmsk.uitm.edu.my

Kamaruzaman Jusoff (Corresponding author)
Yale's Centre for Earth Observation, Environmental Science Centre
New Haven, CT 06511, USA
Tel: 203-432-1384 Email: jusoff.kamaruzaman@yale.edu

Abstract

This paper presents different methods of handling pronunciation variations in Standard Malay (SM) speech recognition. Pronunciation variation can be handled by explicitly modifying the knowledge sources or improving the decoding method. Two types of pronunciation variations are defined, namely, complete or phone changes and partial or sound changes. Complete or phone change means that one phoneme is realized as another phoneme. Meanwhile, a partial or sound change happens when the acoustic realization is ambiguous between two phonemes. Complete or phone changes can be handled by constructing a pronunciation variation dictionary to include alternative pronunciations at the lexical level or dynamically expanding the search space to include those pronunciation variants. Sound or partial changes can be handled by adjusting the acoustic models through sharing or adaptation of the Gaussian mixture components. Experimental results show that the use of a pronunciation variation dictionary and the method of dynamic search space expansion can improve speech recognition performance substantially. The methods of acoustic model refinement were found to be relatively less effective in our experiments.

Keywords: Pronunciation variations, Standard Malay (SM), Complete changes, Partial Changes

1. Introduction

The tremendous growth of technology increased the need of integration of spoken language technologies into our daily applications, providing an easy and natural access to information. These applications are of different nature with different user interfaces. Besides voice enabled Internet portals or tourist information systems, Automatic Speech Recognition (ASR) systems can be used in home user experiences where TV and other appliances could be voice controlled, discarding keyboards or mouse interfaces, or in mobile phones and palm-sized computers for a hands-free and eyes-free manipulation.

Speech is a process used to communicate from a speaker to a listener. Pronunciation relates to speech, and humans have an intuitive feel for pronunciation. For instance, people chuckle when words are mispronounced and notice when foreign accent colors a speaker's pronunciations (Strik and Cucchiarini, 1999). If words were always pronounced in the same way, ASR would be relatively easy. However, for various reasons words are almost always pronounced differently and varied from one speaker to another and from one situation to another. The variability is due to co-articulation, regional accents, speaking rate, speaking style, etc. Pronunciation variation can be further classified into two types: *complete changes* and *partial changes* (Fung *et al*., 2000; Li *et al*., 2000; Saraclar and Khudanpur, 2000; Liu, 2002; Kam, 2003; Liu and Fung, 2003). Complete changes, or phone changes, are the replacement of a phoneme by another alternate phone, such as 'b' being pronounced as 'p'. Partial changes, or sound changes, are the variations within the phoneme such as nasalization, centralization, voiceless and voiced. Both complete changes and partial changes are very common in spontaneous speech. Research in ASR has gradually progressed from isolated words, via connected words and carefully read speech to conversational or spontaneous speech. Although many current application still make use of isolated word recognition example in dictation system, in ASR research the emphasis is now on spontaneous or conversational speech. It is clear that in going from isolated words to conversational or spontaneous speech the amount pronunciation increases. This is because spontaneous speech contains much more phone change (substituted, deleted, and inserted) phenomena and sound change (nasalized, centralized, voiced, voiceless, more rounded, syllabic, pharyngealized, and aspirated) phenomena because of variable speaking rates, moods, emotions, prosodies, co-articulations and so on, even when the speaker is tending to utter in canonical pronunciations (Greenberg, 1999). Other phenomena, such as lengthening, breathing, disfluency, lip smacking, murmuring, coughing, laughing, crying, modal/exclamation, silence, and noise, will also bring difficulties to ASR systems. At the linguistic level, there are a lot of spoken language phenomena, such as repetitions, ellipses, corrections, hesitations, and so on, resulting from the fact that people are often thinking while speaking in daily life. Therefore, since the presence of variation in pronunciation can cause errors in ASR, modeling pronunciation variation is seen as possible way of improving the performance of the current system.

There have been many studies on modeling pronunciation variations for improving ASR performance. They are focused mainly on two problems, namely the prediction of the pronunciation variants, and effective use of pronunciation variation information in the recognition process (Strik and Cucchiarini, 1999). Knowledge-based approaches use findings from linguistic studies, existing pronunciation dictionaries, and phonological rules to predict the pronunciation variations that could be encountered in ASR (Aubert and Dugast, 1995; Kessens *et al*., 1999). Data-driven approaches attempt to discover the pronunciation variants and the underlying rules from acoustic signals. This is done by performing automatic phone recognition and aligning the recognized phone sequences with reference transcriptions to find out the surface forms (Saraçlar *et al*., 2000; Wester, 2003). Some studies used hand-labelled corpora (Riley *et al*., 1999).

The key components of speech recognition system are the acoustic models, the pronunciation lexicon and the language models (Huang *et al*., 2001). The acoustic models are a set of hidden Markov models (HMM) that characterize the statistical variations of input speech. Each HMM represents a specific sub-word unit, e.g. a phoneme. The pronunciation lexicon and the language models are used to define and constrain the ways sub-word units can be concatenated to form words and sentences. They are used to define a search space from which the most likely word string(s) can be determined with a computationally efficient decoding algorithm. Within such a framework, pronunciation variations can be handled by modifying one or more of the knowledge sources or improving the decoding algorithm. Phone changes can be handled by replacing the baseform transcription with surface-form transcriptions, i.e. the actual pronunciations observed. This can be done by either augmenting the baseform lexicon with the additional pronunciation variants (Kessens *et al*., 1999; Liu *et al*., 2000; Byrne *et al*., 2001) or expanding the search space during the decoding process to include those variants (Kam and Lee, 2002). In order to deal with sound changes, pronunciation modeling must be applied at a lower level, for example, on the individual states of a hidden Markov model (HMM) (Saraçlar *et al*., 2000). In general, acoustic models are trained solely with baseform transcriptions. It is assumed that all training utterances follow exactly the canonical pronunciations. This convenient, but apparently unrealistic, assumption renders the acoustic models inadequate in representing the variations of speech sounds. To alleviate this problem, various methods of acoustic model refinement were proposed (Saraçlar *et al*., 2000; Venkataramani and Byrne, 2001; Liu, 2002).

In this paper, the pronunciation variations in spontaneous Standard Malay (SM) speech are studied. The linguistic and acoustic properties of spoken SM language are considered in the analysis of pronunciation variations and, subsequently, the design of pronunciation modeling techniques for the speech database. There are 500 million people that speak Bahasa Melayu or Bahasa Malaysia and it is the official language in Malaysia, Indonesia and Brunei. This language is part of Austronesian language and it is agglutinative in nature, that is the words in Bahasa Melayu are formed by joining syllables. The term "Standard Malay" (SM) is a term that is basically accepted by the speech community to be the norm or the prestige dialect, which is also the official language in Malaysia. It is widely believed that the so-called "Standard Malay" is based on the Johor-Riau Malay (JM) dialect, mainly spoken in the southern part of the peninsular Malaysia. There are other three dialects, namely Kelantan Malay (KM), Ulu Muar Malay (UMM) and Langkawi Malay (LM) which are soken in the different parts of peninsular Malaysia (Teoh, 1994).

There are some common features between the Malay language and English language. Firstly, Malay language is a phonetic language and it is written in Roman characters. Secondly, all syllables in the Malay language are pronounced almost equally and it is thus, considered as a non-tonal language. In general, there are six (6) main vowels and 29 consonants in SM. SM have a total of nineteen of the consonants, where /m/, /n/, /f/, /l/, /s/ and /y/ are pronounced almost the same way as in English. In Malay language, the syllabic structure is well-defined and can be unambiguously derived from a phone string. The basic syllable structure of the Malay language is generated by an ordered series of three syllabication rules. The linguists claimed that Malay is a Type III language (Teoh, 1994), namely Consonant-Vowel (CV) and Consonant-Vowel-Consonant (CVC) are the most common and they can be found almost in every Malay primary word. Based on the CV(C) structure, coda is optional for the syllable in Malay language and open syllable are commonly found. For Standard Malay language, the alphabets in a word itself is good enough to identify its pronunciation. However not all words, can be pronounced exactly as it written. The structure of the syllable (open or closed) and the position of the syllable (initial, middle or final) control the distribution of the SM vocalic segments. For instance the vowels /e/ and /o/ does not occur in open syllables. If this vowels occur, they will be removed by the final deletion rule (example: *"bazir"* = /b.a.z.e./ (waste)). SM also does not have /a/ occurring in the final position except in borrowed words such as a *"baba"* /b. a. b. a./ (Malaysia Straits born Chinese) and *"lawa"* /l. a. w. a./ (attractive, beautiful). This types of pronunciation variations can be classified as complete or phone changes.

Meanwhile the SM consonants are non-syllabic, the SM vowels are syllabic. Oral vowels can also be nasalized vowel, where vowels are followed by nasal sound. For example the SM word *"minggu"* (week) /m. i. ng. u/, where the vowel "i" is nasalized as it is followed by the velar nasal /n/. The length of the vowel is not distinctive and it is not a feature that differentiates one vowel phoneme from another. In addition to the vowel depending on the context in which they occur, vowels can be long or short, and these pronunciation variations could be classified as partial or sound changes. There are lots more word variations if we compare between native and non-native speakers pronunciations because not all people that speak Standard Malay use the same pronunciation (El-Imam *et al*., 2000) and it more variations during spontaneous speech.

2. Development of speech database

In building the Standard Malay speech database, the selections of utterances are derived from Buletin Utama TV3 Broadcast News that contains about 550 utterances for four hours news. All recognition experiments described in this paper use the Hidden Markov Toolkit (HTK) version 3.2 (Young *et al*., 2001). We trained the recognition model based on syllables that formed by concatenating three types of phonological units: the Initial, the Middle and the Final that represented as a sequence of (SM) characters as shows in Table 1 and Table 2. For the purpose of this study, we focus on the Initial and Final (IF) representation. As a recognition feature, we extract 12 mel-frequency cepstral coefficients (MFCCs) with a logarithmic energy for every 10 ms analysis frame, and concatenate their first and second derivatives to obtain a 39-dimensional feature vector. During training and testing, we apply cepstral mean normalization and energy normalization to the feature vectors. The whole training procedure is divided into two stages, where monophone and triphone stages should be applied. In each stage, there are always two steps, which are repeated iteratively by estimation and realignment. The process begins with the training of the monophone models, followed by training of the triphone models. The acoustic models are based on 3-state left to right, context-dependent, 4-mixture, and cross-word triphone models, trained using the HTK toolkit (Young *et al*., 2001).

3. Recognition modeling and implementation

3.1Modeling pronunciation variation for complete/phone changes

The pronunciation lexicon used in the baseline system provides only the baseform pronunciation for each of the word entries. In real speech, the baseform pronunciations are realized differently, depending on the speakers, speaking styles, etc. A pronunciation model (PM) is a descriptive and predictive model by which the surface-form pronunciation(s) can be derived from the baseform one. There have been three different types of models proposed by previous studies such as the phonological rules for generating pronunciation variations (Wester, 2003; Kessens *et al*., 2003), a pronunciation variation dictionary (PVD) that explicitly lists alternative pronunciations (Aubert and Dugast, 1995; Kessens *et al*. 1999; Liu *et al*., 2000], and the statistical decision trees that predict pronunciation variations according to phonetic context (Riley *et al*., 1999; Fosler-Lussier, 1999; Saraçlar *et al*., 2000].

In this study, two different approaches to handling phone changes in Standard Malay ASR are formulated and evaluated. The first approach uses a probabilistic PVD to augment the baseform lexicon. This is a straightforward and commonly used method that has been proven effective for various tasks and languages (Strik and Cucchiarini, 1999). In the second approach, pronunciation variation information is introduced during the decoding process. Decision tree based PMs are used to dynamically expand the search space. In (Saraçlar *et al*., 2000) a similar idea was presented. Decision tree based PMs were applied to a word lattice to construct a recognition network that includes surface-form realizations.

In the first approach, the PVD, each word can have multiple pronunciations, each being assigned a word-level variation probability (VP). The PVD can be obtained from the IF confusion matrix. The word-level VP is given by multiplying the phone-level VPs of all the individual phonemes in the surface-form pronunciation. With the use of the PVD, the goal of speech recognition is essentially to search for the most probable word sequence by considering all possible surface-form realizations. This can be made as

$$W^* = \arg\max_{W} P(O \mid B)P(B \mid W)P(W) \tag{1}$$

where $P(O|B)$ and $P(W)$ are referred to as the (sub-word level) acoustic models and the language models, respectively. $P(B|W)$ is given by a pronunciation lexicon. Further Eq. (1) can be modifying as

$$W^* = \arg\max_{W,k} P(O|S_{W,k})P(S_{W,k}|W)P(W) \tag{2}$$

where $S_{W,k}$ denotes one of the surface-forms realizations of W. $P(S_{W,k}|W)$ are obtained from the word-level VPs.

The PVD includes both context-independent and context-dependent phone changes. Since each word is treated individually, the phonetic context being considered is limited to within the word. To deal with cross-word context-dependent phone changes, we propose applying pronunciation models at the decoding level. Our baseline system uses a one-pass search algorithm (Choi, 2001).

In the second approach, a context-dependent pronunciation model is needed to predict the surface-form phoneme given the baseform phoneme and its context. It is implemented using the decision tree clustering technique, following the approaches described in (Riley *et al*., 1999; Fosler-Lussier, 1999). Each baseform phoneme is described using a decision tree. Given a baseform phoneme, as well as its left context (the right context is not available in a forward Viterbi search), the respective decision-tree pronunciation model (DTPM) gives all possible surface-form realizations and their corresponding VPs (Kam and Lee, 2002). Like the confusion matrix, the DTPM is trained with the phoneme recognition outputs for the HTK toolkit training utterances. The training involves an optimization process by which the surface-form phonemes are clustered based on phonetic context. At a particular node of the tree, a set of "yes/no" questions about the phonetic context are evaluated. Each question leads to a different partition of the training data. The question that minimizes the overall conditional entropy of the surface-form realizations is selected for that node. The node-splitting process stops when there are too few training data (Kam, 2003).

3.2 Modeling pronunciation variation for partial/sound changes

Unlike phone changes, a sound change cannot be described as a simple substitution of one phoneme for another. It is regarded as a partial change from the baseform phoneme to a surface-form phoneme (Liu and Fung, 2003). Our approaches presented below attempt to refine the acoustic models to handle the acoustic variation caused by sound changes. The acoustic models are continuous-density HMMs. The output probability density function (pdf) at each HMM state is a mixture of Gaussian distributions. The use of multiple mixture components is intended to describe complex acoustic variabilities. In this study, we investigate both the sharing and adaptation of the acoustic model parameters at the mixture level (Kam *et al*., 2003). In the first approach, sharing of mixture components is applied and the states of the baseform and surface-form models are aligned. It is assumed that both models have the same number of states. Then, state *j* of the baseform model is aligned with state *j* of the surface-form model. Consider a baseform phoneme *B*. The output pdf at state *j* is given as

$$b_j(o_t) = \sum_{m=1}^{M} w_{jm} N(o_t; \mu_{jm}, \Sigma_{jm}) \tag{3}$$

where M is the number of Gaussian mixture components, and w_{jm} is the weight for the m th mixture component. The baseform output pdf can be modified to include the contributions from the surface-form states

$$bj'(o_t) = VP(B,B).b_j(o_t) + \sum_{n=1}^{N} VP(S_n,B).qs_n, j(o_t)$$

$$S_n \neq B \tag{4}$$

where S_n denotes the n th surface-form of B, N is the total number of surface-forms, $VP(S_n,B)$ is the variation probability of S_n with respect to baseform B, and $qs_n, j(o_t)$ denotes the output pdf of state j of the n th surface-form model.

The number of mixture components in the resultant baseform model depends on N . More surface-form pronunciations bring in more mixture components to the modified baseform state. As the number of mixture components is changed, re-estimation of mixture weights is required. Although sharing mixture components yields an acoustically richer model, it also greatly increases the model size for which more memory space and higher computation complexities are required. Moreover, if the baseform and surface-form mixture components are very similar, including them all in the modified baseform is unnecessarily superfluous.

For the second approach, we propose to refine the baseform acoustic models through parameters adaptation. The total number of model parameters remains unchanged. Like in the approach of mixture sharing, the states of the baseform and surface-form models are aligned. The surface-forms are generated from the IF confusion matrix. Consider the aligned states of the baseform phoneme B and one of its surface-forms S. Let $m_B(i)$ and $m_S(j)$ denote the *i*th mixture component in the baseform state and the *j*th mixture component in the surface-form state, respectively, where

$i, j = 1,2,..., M$. The distances between all pairs $(m_B(i), m_S(j))$ are computed. Then each surface-form component is paired up with the nearest baseform component. That is, for each $m_S(j)$, we find

$$\hat{i} = \arg\min_{m_B(i)} d(m_B(i), m_S(j)) \tag{5}$$

The "distance" between two Gaussian distributions is calculated using the Kullback-Leibler divergence (KLD) (Myrvoll and Soong, 2003). Given two multivariate Gaussian distributions f and g , the symmetric KLD has the following closed form

$$d(f,g) = \frac{1}{2} trace\{(\Sigma_f^{-1} + \Sigma_g^{-1})(\mu_f - \mu_g)(\mu_f - \mu_g)^T + \Sigma_f \Sigma_g^{-1} + \Sigma_g \Sigma_f^{-1} - 2I\} \tag{6}$$

where μ and Σ denote the mean vectors and the covariance matrices of the two distributions, respectively, and I is the identity matrix.

As a result, for this pair of baseform and surface-form states, each Gaussian component $m_B(i)$ is associated with k surface-form components, as illustrated in Figure 1. The centroid of these k components is computed. If the baseform B has n surface forms, there will be n such centroids. These surface-form centroids and the corresponding baseform component are weighted with the VP, and together produce a new centroid that is taken as the adapted baseform component. In this way, the adapted model is expected to shift towards the surface-form phonemes. The extent of such a shift depends on the VP. The mean and covariance of the centroid of k weighted Gaussian components can be found by minimizing the following weighted divergence

$$\{\mu_c', \Sigma_c'\} = \arg\min_{\mu_c, \Sigma_c} \sum_{n=1}^{k} a_n d(f_c, f_n) \tag{7}$$

where f_n denotes the nth component and a_n is the respective weighting coefficient. Assuming diagonal covariances, the weighted centroid is given as (Myrvoll and Soong, 2003)

$$\mu_c'(i) = \frac{\Sigma_{n=1}^{k} a_n (\Sigma_c^{-1}(i) + \Sigma_n^{-1}(i))\mu_n(i)}{\Sigma_{n=1}^{k} a_n (\Sigma_c^{-1}(i) + \Sigma_n^{-1}(i)}$$

$$\Sigma_c'(i) = \sqrt{\frac{\Sigma_{n=1}^{k} a_n \left[\Sigma_n(i) + (\mu_c(i) - \mu_n(i))^2\right]}{\Sigma_{n=1}^{k} a_n \Sigma_n^{-1}(i)}} \tag{8}$$

4. Results and discussions

4.1Pronunciation variation modeling for complete/phone changes

The recognition results with the use of PVDs that are constructed with different values of the VP threshold are shows in Table 3. The baseline system uses the basic pronunciation lexicon that contains 451 words. The size of the PVD increases as the VP threshold decreases. It is obvious that the introduction of pronunciation variants improves recognition performance. The best performance is attained with a VP threshold of 0.05. In this case, the PVD contains 568 pronunciations for the 451 words. With a very small value for the VP threshold, e.g. 0.02, the recognition performance is not good because there are too many pronunciation variants being included and some of them do not really represent pronunciation variation.

The recognition result attained by using the DTPM for dynamic search space expansion is show in Table 4. It appears that this approach is as effective as the PVD. Unlike the results for the PVD, the performance with a VP threshold of 0.2 is better than that with a threshold of 0.05. This means that the predictions made by the DTPM should be pruned more stringently than the IF confusion matrix. Because of its context-dependent nature, the DTPM has relatively less training data, and the variation probabilities cannot be reliably estimated. It is preferable not to include those unreliably predicted pronunciation variants.

By analyzing the recognition results in detail, it is observed that many errors are corrected by allowing the following pronunciation variations:

Initials: [b]→[p] or [m], [d]→[l] or [t], [g]→ [k], [s]→ [t] ,[t]→ [d]

Finals: [ang]→[an], [sy]→[sa] or [su], [ng]→[m] (syllabic nasal)

These observations match well with the findings in sociolinguistic studies on Standard Malay phonology.

4.2 Pronunciation variation modeling for partial/sound changes

The recognition results attained with the two methods of acoustic model refinement as shows in Table 5. The VP threshold for surface-form prediction is set at 0.05. Apparently, both approaches improve recognition performance. The sharing of mixture components seems to be more effective than adaptation. However, this is at the cost of a substantial increase in model complexity. The baseline acoustic models have a total of 2,251 Gaussian components. The adaptation approach retains the same number of Gaussian components. The models obtained with the sharing approach have 2,620 components, 17% more than the baseline. If we use an equal number of components in the baseline acoustic models, the baseline word error rate will be reduced to 14.34%, and the benefit of sharing mixture components is only marginal.

With the adaptation approach, the baseform pdf is shifted towards the corresponding surface forms. If a surface-form pdf is far away from the baseform one, the extent of the modification will be substantial and, consequently, the modified pdf may fail to model the original baseform. On the other hand, the sharing approach has the problem of undesirably including redundant components in the baseform models. Thus we combine these two approaches. The idea is to perform adaptation using the surface-form components that are close to the baseform, and at the same time, to use those relatively distant components for sharing.

The values of the KLD between the baseform pdf and the nearest surface-form pdf have been analyzed. As illustrative examples, the histograms of the KLD at different states between [oi] (baseform) and [o] (surface form), and between [oi] and [ou], are shown as in Figure 2. There are two main types of KLD distributions: 1) concentration around small values (e.g., states 1 and 2 of the pair "[oi]→[o]"), and 2) a wide range of values (e.g., states 3 to 5 of the pair "[oi]→[ou]"). A small KLD means that the mixture components of the baseform and surface forms are similar. In this case, the baseform components adapt to the surface form. In the case of a widely distributed KLD, the surface-form components should not be used to adapt the baseform components, but rather should be kept along with the modified baseform model in order to explicitly characterize irregular pronunciations. In this way, a combined approach to baseform model refinement is formulated.

Despite the good intentions, the combined use of sharing and adaptation does not lead to favorable experimental results. With a total of 2,441 mixture components in the refined acoustic models, the word error rate is 14.57%. The baseline performance is 14.93% with the same model complexity.

5. Conclusion

In this study, we have classified pronunciation variations into complete/phone changes and partial/sound changes. However, these are not well defined classifications, especially for the partial/sound changes. There is not a clear boundary that separates a phoneme substitution (phone change) from a phoneme modification (sound change). This may partially explain why the proposed techniques of handling sound change are not as effective as the methods for handling phone change.

The use of a PVD is intuitive and straightforward in implementation. It can reduce the word error rate noticeably. When constructing a PVD, the value of the VP threshold needs to be carefully determined. While a tight threshold obviously does not show any effect, a lax control of the PVD size leads to not only a long recognition time but also performance degradation. The method of dynamic search space expansion during decoding can bring about the same degree of performance improvement as the PVD. However, the training of context-dependent pronunciation prediction models requires a large amount of data.

The methods of acoustic model refinement do not improve recognition performance as much as we expected. Similar effect can be achieved by using more mixture components. Indeed, more mixture components can describe more complex acoustic variations, which include the variations caused by alternative pronunciations. The sharing of mixture components is equivalent to having more mixture components right at the beginning of acoustic models training. Adaptation of mixture components is not as effective as increasing the number of mixture components.

For any of the above methods to be effective, the accurate and efficient acquisition of pronunciation variation information is most critical. Manual labeling is impractical. Automatic detection of pronunciation variations is still an open problem.

References

Aubert, X., and C. Dugast, "Improved acoustic-phonetic modeling in Philips' dictation system by handling liaisons and multiple pronunciations," In *Proceedings of 1995 European Conference on Speech Communication and Technology*, pp.767 – 770.

Byrne, W., V. Venkataramani, T. Kamm, T.F. Zheng, Z. Song, P. Fung, Y. Liu and U. Ruhi, "Automatic generation of pronunciation lexicons for Mandarin spontaneous speech," In *Proceedings of the 2001 International Conference on Acoustics, Speech and Signal Processing*, l.1, pp.569 – 572.

Choi, W.N., *An Efficient Decoding Method for Continuous Speech Recognition Based on a Tree-Structured Lexicon*, MPhil Thesis, The Chinese University of Hong Kong, 2001.

El-Imam, Y.A., and Don, Z.M., "Text-to-Speech Conversion of Standard Malay", International Journal of Speech Technology 3, Kluwer Academic Publishers, pp. 129-146, 2000.

Fung, P., Byrne, W., Zheng, F., Kamm, T., Liu, Y., Song, Z., et al. (2000). Pronunciation modeling of mandarin casual speech. *2000 Johns Hopkins Summer Workshop*.

Fosler-Lussier, E., "Multi-level decision trees for static and dynamic pronunciation models," In *Proceedings of 1999 European Conference on Speech Communication and Technology*, pp.463 – 466.

Greenberg, S., "Speaking in shorthand – a syllable centric perspective for understanding pronunciation variation", *Speech Communication* 29 (2-4), pp. 159-176, 1999.

Huang, X., A. Acero, and H.W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*, Prentice Hall PTR., 2001.

Kam, P., *Pronunciation Modeling for Cantonese Speech Recognition*, MPhil Thesis, The Chinese University of Hong Kong, 2003.

Kam, P., and T. Lee, "Modeling pronunciation variation for Cantonese speech recognition," In *Proceedings of ISCA ITR-Workshop on Pronunciation Modeling and Lexicon Adaptation* 2002, pp.12-17.

Kam, P., T. Lee and F. Soong, "Modeling Cantonese pronunciation variation by acoustic model refinement," In *Proceedings of 2003 European Conference on Speech Communication and Technology*, pp.1477 – 1480.

Kessens, J.M., M. Wester and H. Strik, "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation," *Speech Communication*, 29, pp.193 – 207, 1999.

Kessens, J.M., C. Cucchiarini and H. Strik, "A data driven method for modeling pronunciation variation," *Speech Communication*, 40, pp.517 – 534, 2003.

Li, A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., et al. (2000). CASS: A Phonetically Transcribed Corpus of Mandarin Spontaneous Speech. *Sixth International Conference on Spoken Language Processing*.

Liu, M., B. Xu, T. Huang, Y. Deng and C. Li, "Mandarin accent adaptation based on context-independent/context-dependent pronunciation modeling," In *Proceedings of the 2000 International Conference on Acoustics, Speech and Signal Processing*, 2, pp.1025-1028.

Liu, Y., *Pronunciation Modeling for Spontaneous Mandarin Speech Recognition*, PhD Thesis, The Hong Kong University of Science and Technology, 2002.

Liu, Y. and P. Fung, "Modeling partial pronunciation variations for spontaneous Mandarin speech recognition," *Computer Speech and Language*, 17, 2003, pp.357 – 379.

Myrvoll, T.A. and F. Soong, "Optimal clustering of multivariate normal distributions using divergence and its application to HMM adaptation", In *Proceedings of the 2003 International Conference on Acoustics, Speech and Signal Processing*, 1, pp.552 - 555.

Riley, M., W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, Saraclar, M., Wooters, C., Zavaliangkos, G., 1999. Stochastic pronunciation modeling from hand-labelled phonetic corpora. Speech Communication 29 (2-4) 209-224.

Saraçlar, C. Wooters and G. Zavaliagkos, "Stochastic pronunciation modeling from hand-labelled phonetic corpora," *Speech Communication*, 29, 1999, pp.209 – 224.

Saraçlar, M. and S. Khudanpur, "Pronunciation ambiguity vs. pronunciation variability in speech recognition," In *Proceedings of the 2000 International Conference on Acoustics,Speech and Signal Processing*, 3, pp.1679-1682.

Saraçlar, M., H. Nock and S. Khudanpur, "Pronunciation modeling by sharing Gaussian densities across phonetic models," *Computer Speech and Language*, 14, 2000, pp.137 – 160.

Strik, H. and C. Cucchiarini, "Modeling pronunciation variation for ASR: a survey of the literature," *Speech Communication*, 29, 1999, pp.255 – 246.

Teoh, B. S. 1994. The Sound system of Malay revisited. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Venkataramani, V. and W. Byrne, "MLLR adaptation techniques for pronunciation modeling," In *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding* 2001, CD-ROM.

Wester, M., "Pronunciation modeling for ASR – knowledge-based and data-derived methods," *Computer Speech and Language*, 17, 2003, pp.69 – 85.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., 2001. The HTK Book. Cambridge University Engineering Department.

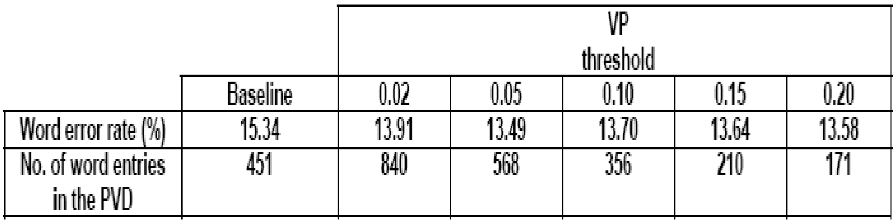
Table 1. Classification of vowel sounds in Standard Malay language

	Initial	**Middle**	**Final**
High	i		u
Medium	e	at	o
Low		a	

Table 2. Consonants in Standard Malay except those in brackets are loaned consonants

Manner of Articulation	**Place of articulation**					
	Labial	**Alveolar**	**Palate-alveoral**	**Palatal**	**Velar**	**Glottal**
Plosive-Voiceless	p	t			k	
Plosive-Voiced	b	d			g	
Fricative-Voiceless	(f)	s			(x)	h
Fricative-Voiced	(v)	(z)				
Affricate-Voiceless			c			
Affricate-Voiced			j			
Nasal	(m)	n		ny	ng, nx	
Roll		r				
Lateral		l				
Semivowel	w			y		

Table 3. Recognition results of using a PVD with different VP thresholds



		VP threshold				
	Baseline	0.02	0.05	0.10	0.15	0.20
Word error rate (%)	15.34	13.91	13.49	13.70	13.64	13.58
No. of word entries in the PVD	451	840	568	356	210	171

Table 4. Recognition results by dynamic search space expansion

		VP	threshold
	Baseline	0.05	0.2
Word error rate (%)	15.34	13.53	13.27

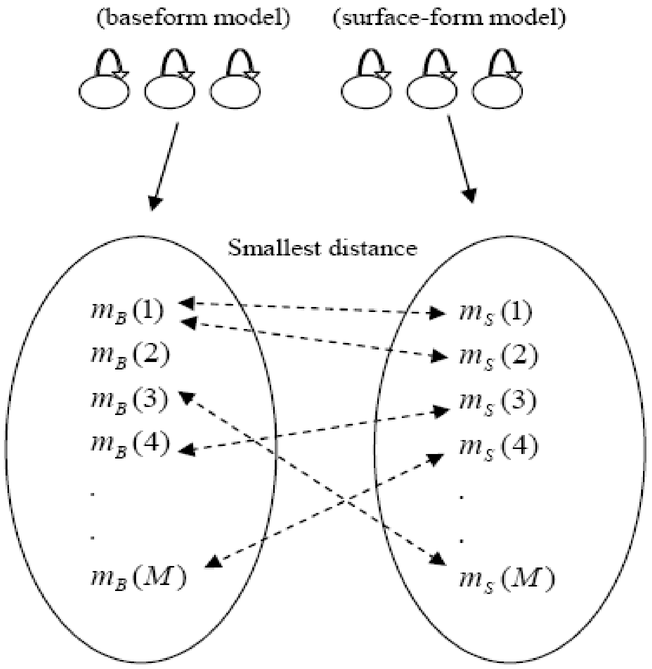


Figure 1. Mapping between baseform and surface-form mixture components

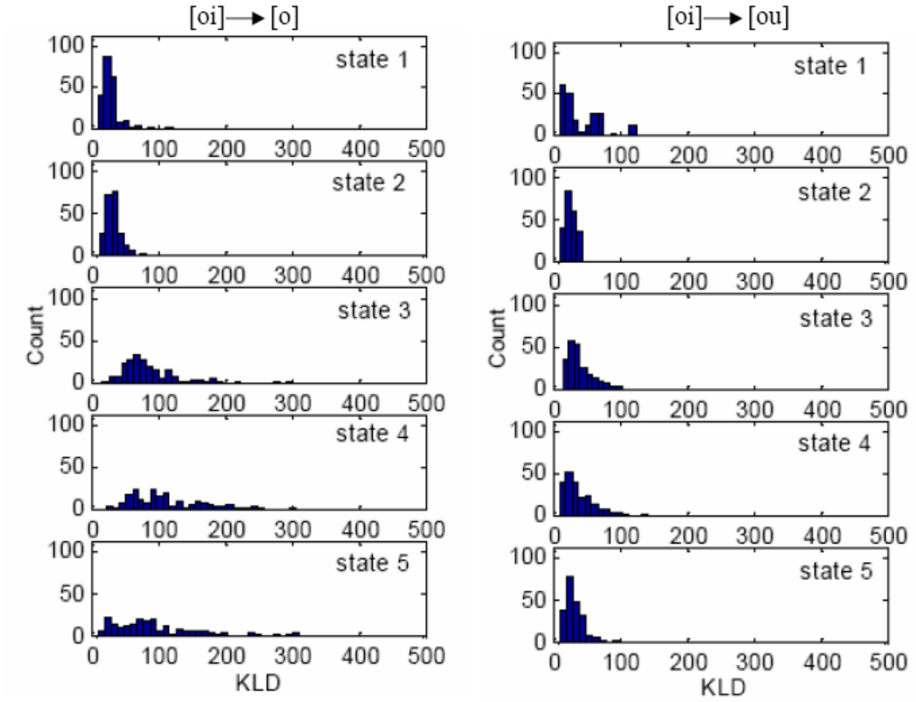


Figure 2. KLD distributions for variation pairs [oi]→[o] and [oi]→[ou]