

PRONUNCIATION MODELING FOR SPEECH TECHNOLOGY

Torbjørn Svendsen

Department of Electronics and Telecommunications
NTNU
N-7491 Trondheim, NORWAY
torbjorn@iet.ntnu.no

ABSTRACT

Written text is based on an orthographic representation of words, i.e. linear sequences of letters. Modern speech technology (automatic speech recognition and text-to-speech synthesis) is based on phonetic units representing realization of sounds. A mapping between the orthographic form and phonetic forms representing the pronunciation is thus required. This may be obtained by creating pronunciation lexica and/or rule-based systems for grapheme-to-phoneme conversion. Traditionally, this mapping has been obtained manually, based on phonetic and linguistic knowledge. This approach has a number of drawbacks: i) the pronunciations represent typical pronunciations and will have a limited capacity for describing pronunciation variation due to speaking style and dialectal/accents variations; ii) if multiple pronunciation variants are included, it does not indicate which variants are more significant for the specific application; iii) the description is based on phonetic knowledge and does not take into account that the units used in speech technology may deviate from the phonetic interpretation; and iv) the description is limited to units with a linguistic interpretation. The paper will present and discuss methods for modeling pronunciation and pronunciation variation specifically for applications in speech technology.

1. INTRODUCTION

Automatic speech recognition (ASR) is basically about establishing a mapping from a spoken utterance, the speech waveform, to its orthographic representation in the form of text. Conversely, text-to-speech synthesis (TTS) concerns converting a written text to a speech waveform. In both cases, the mapping normally contains an abstract description of the pronunciation in terms of some basic unit. In most cases, this basic unit is the phone, or some derivation thereof. If we disregard coarticulation effects at word boundaries, a representation of a textual utterance (e.g. a sentence) can be created by concatenating the phonetic representations of the words contained in the sentence. Adding predicted prosodic information (phone

duration, loudness and F0) this constitutes the basic information necessary to convert the written sentence to speech. Similarly, in ASR the decoding of the spoken utterance consists of finding the most likely phone sequence that corresponds to a legal word sequence. It is obviously of some importance for the performance of ASR and TTS that the predicted pronunciation corresponds well with the way words and utterances are actually pronounced.

Two important issues immediately present themselves:

- Presuming that a “typical” pronunciation is desired, how can the typical pronunciation best be obtained in terms of optimal performance of the speech technology application?
- It is obvious that the pronunciation will vary depending on the speaker’s accent or dialect, and that the speaking style and emotional state of the speaker will influence strongly on pronunciation. How can we best obtain a description of pronunciation variation due to e.g. speaker accent/dialect and speaking style?

The paper is organized as follows: In section 2, an overview of methods for pronunciation modeling is presented. Section 3 gives selected examples of pronunciation modeling methods and applications.

2. PRONUNCIATION MODELING

The basic method for describing the pronunciation of words is through the use of a pronunciation lexicon. In the pronunciation lexicon, a lexical entry will contain the keyword (e.g. orthographic word), optional grammatical information and pronunciation information. The pronunciation information will consist of one or more *baseforms*. A baseform is a linear sequence of basic units (e.g. phones) describing a pronunciation of the lexical keyword. If there are more than one baseform for a word, the baseforms describe allowed pronunciation variants of the word. For some applications, e.g. TTS, the baseform information is augmented with information about syllable boundaries, syllable stress etc. In ASR, grammatical information can be embedded into the keyword, e.g. OBJECT_n to signify the noun vs. OBJECT_v to signify the verb.

Some general issues need to be considered before constructing a pronunciation lexicon. The question of whether orthographic and pronunciation ambiguities will influence the application, and if so, how to do it, needs to be addressed. Examples of ambiguities are homographs (words that are identically spelled, but have different meanings and pronunciations), homonyms (words with identical orthography and pronunciation, but different meanings) and homophones (words with identical pronunciations, but different meanings and possibly different spellings). Secondly, the issue of whether or not to include pronunciation variants in the lexicon must be resolved.

The traditional method of generating a pronunciation lexicon is by manual construction. This is typically done by a skilled phonetician (or derived from an existing pronunciation dictionary originally constructed in this manner). Apart from being costly and labor intensive, this method has a number of potential drawbacks: i) the pronunciations represent typical pronunciations and will have a limited capacity for describing pronunciation variation due to speaking style and dialectal/accents variations; ii) if multiple pronunciation variants are included, it does not indicate which variants are more significant for the specific application; iii) the description is based on phonetic knowledge and does not take into account that the units used in speech technology may deviate from the phonetic interpretation; and iv) the description is limited to units with a linguistic interpretation. These issues are addressed by pronunciation modeling.

There have been a multitude of proposals of how to model pronunciation. The approaches can generally be divided into two main categories, *knowledge-based* and *data-driven* [1]. A further sub-division into *direct* and *indirect* methods can be applied [2] indicating hybrids, e.g., indirect knowledge-based modeling uses data-driven methods to improve a basic knowledge-based model.

2.1. Knowledge-based pronunciation modeling

In knowledge-based pronunciation modeling, the lexicon is generated from a set of pronunciation rules. The rule set constitutes a grapheme-to-phoneme mapping, and is based on explicit rules derived from linguistic knowledge. E.g. the rule that a word initial “ch” letter combination should always be pronounced /ch/ except when followed by an “r” can be expressed by the rules:

```
(# [ch] r => k)
(# [ch] => ch)
```

provided that the rule strategy chooses the first matching rule and the rules are organized as shown above [3]. Rules can be built by hand, or extracted from an existing large pronunciation lexicon using e.g. Classification and Regression Trees (CART).

Pronunciation rules can be used for generating canonic lexica or, with an appropriate rule-set and rule selection

strategy, also for generating lexica containing pronunciation variants. Rules can also be applied to express cross-word effects when constructing pronunciation strings for an utterance consisting of several words.

Pure knowledge-based pronunciation modeling does not supply a remedy to the drawbacks of traditional lexicon construction. In particular, general purpose rules are not well suited to model conversational and/or accented/dialectal speech. A study of manual phonemic transcriptions of the conversational Switchboard corpus has e.g. revealed that only a third of the labeled word pronunciations were found in the Pronlex dictionary [4]. At least for speech recognition, efforts in pronunciation modeling today tend to at least include data-driven methodologies.

2.2 Indirect knowledge-based pronunciation modeling

There are two obvious methods of improving knowledge-based pronunciation modeling based on data. The first method requires that the rule-set produce a pronunciation lexicon that contains a large number of pronunciation variants. Using available speech data, the variant lexicon can be pruned to include the most frequent pronunciation variants. The pruning can e.g. be based on ASR performance or the frequency count of variants obtained from forced alignment employing the allowed variants of each word in an utterance.

Alternatively, the speech data can be utilized to augment a canonical lexicon with variants. In [5], a standard canonic lexicon constitute the knowledge-based baseline. Using hand-labeled corpora, and a “reference” transcription built using a variant pronunciation lexicon aided by forced alignment to select the best matching pronunciation variant, decision trees were built which predict realized phoneme-to-phone mappings. The dictionary was then expanded using the decision trees to predict the most likely pronunciation variants.

2.3 Direct data-driven pronunciation modeling

In data-driven pronunciation modeling, speech databases are used in order to learn the pronunciation (and pronunciation variants). The advantage of data-driven pronunciation is that the learning algorithms can be tied directly to the unit representation in the speech technology application, so that we can avoid mismatches between the lexical modeling and the modeling of the units. For pronunciation variation modeling, data-driven methods will also simplify estimation of variant probabilities if these are desired. The disadvantage is of course limitations in generalization capabilities, and the need for representative databases. The data-driven approaches will be limited to learning pronunciations that are represented in the available speech material, and will require sufficient data material to be able to make good estimates.

Direct data-driven modeling will only be able to model the pronunciations of words that are actually present in the training data. Basically, this type of pronunciation modeling will aim to select the baseform(s) that maximize some objective function on the training data. A simple approach is to transcribe all training occurrences of a word using a phone recognizer. The resulting set of phone sequences constitutes the baseform candidates. The baseforms to be used in the lexicon can be found by evaluating an objective function for each candidate on the training data and selecting the most successful candidate(s). The objective function can e.g. be the likelihood of the training data given the baseform, or the inverse of the edit distance (Levenshtein distance) between the candidate and all other (or k nearest) candidates.

2.4 Indirect data-driven pronunciation modeling

Directly modeling the pronunciation of each vocabulary word from training data requires all vocabulary words to be well represented in the training data. An alternative approach is to use the training data to derive pronunciation rules that in turn can be applied to generate one or more baseforms for any vocabulary word. The advantage of this approach is the ability to generalize, i.e. that the pronunciation of a word can be predicted even if that specific word is not present in the training data. Although indirect data-driven approaches in principle can be applied for finding canonical pronunciation lexica, they are most often used for deriving pronunciation variants.

A typical approach would require a reference transcription to be produced for each training utterance, e.g. from a TTS front end or some existing canonic pronunciation lexicon. An alternative transcription is then generated, e.g. by using a phone recognizer and aligned with the reference. The transcription pairs will exhibit differences, i.e. insertions, deletions and substitutions. These differences can be described by rules. The rule set that explains all the differences between the set of transcription pairs will be too large, and will allow too many mappings to be useful, so a reduced set of rules is required. Such a set can be generated e.g. by phonetic decision trees, or by explicit assessment and pruning of the rule candidates. The resulting rule set can then be applied to generate candidates for a variant pronunciation lexicon. However, the rules will tend to allow significantly more variants than are practically useful in the lexicon. Thus, the variants need to be pruned by being evaluated on the training data.

3. EXAMPLES OF DATA-DRIVEN MODELING

A large number of approaches to pronunciation modeling has been proposed in the literature, most with application to automatic speech recognition. In this section, some

selected methods for pronunciation modeling are briefly described.

3.1 ML-based pronunciation modeling

HMM-based speech recognizers are traditionally trained by estimating model parameters that maximize the likelihood of the training data. It would be consistent with the training procedure and with the recognition criterion to apply the same criterion to the creation of the pronunciation lexicon. Assuming that for each word, W_v , we have a set of N_v training utterances $\{X_v^{(n)}\}$. The optimal baseform B'_v in the ML sense is then found as

$$B'_v = \operatorname{argmax}_B \left\{ \prod_{n=1}^{N_v} p\left(X_v^{(n)} | B\right) \right\} \quad (1)$$

where B is the set of all allowable baseforms. B can be constructed using a grammar. A null grammar will e.g. allow any sequence of phones, of any length. Applying a phone bigram model will weight the baseform candidates with their respective probabilities, favoring baseforms containing frequently occurring phone combinations.

If the number of training utterances $N_v=1$, the optimal baseform is found by using a simple phone recognizer. However, the number of training utterances needs to be large enough to produce representative word models, yielding a non-trivial optimization problem. In principle, the modified tree-trellis algorithm [6] guarantees finding the optimal baseform. In practice, this algorithm can suffer from fan-out problems in the backwards tree search, particularly if the training utterances exhibit large variations. This problem can be reduced or eliminated by employing sub-optimal procedures. Preselection of a subset of the training tokens expected to be less heterogeneous will reduce the training set variability and thus reduce the risk of fan-out problems [6]. Alternatively, one can reduce the search space of baseforms by constructing a set of baseform candidates that is restricted to the phone sequences produced by N -best phone recognition of the training words (see e.g. [11]). The sub-optimal procedures do not impair the performance of the resulting lexicon for recognition, and avoids the convergence problems for the optimization procedure. In [7] experiments were performed on the Resource Management task. This task has a 991 word vocabulary. In the training corpus, 605 vocabulary words had sufficiently many repetitions to be optimized. The pronunciations of the remaining 386 vocabulary words were taken from the original pronunciation dictionary. The lexicon optimization resulted in an error rate reduction of 7-14% when using context independent acoustic models [7].

3.2 ML-based pronunciation variation modeling

The method described above is in principle capable of finding the N best baseforms for a given utterance. However, those baseforms do not describe pronunciation

variations well, they are optimized for describing the entire training set for that utterance. When modeling pronunciation variations, it is desirable to initially identify clusters of the training set, where each cluster contains examples of typical pronunciations of the utterance. Also, not all utterances will exhibit the same amount of pronunciation variation. Thus, the number of pronunciation variants for each word should depend on the relative amount of variation in the pronunciation of that word.

Based on the criteria above, a criterion for modeling pronunciation variation can be formulated. Assume that the pronunciation lexicon L_v for a vocabulary of size V contains a set of baseforms $L_v = \{B_v^{(i)}\}$ where $v=1, \dots, V$ is the index of the word and $i=1, \dots, I_v$ is the index of the pronunciation variant of that word. Assuming that we wish to limit the total size of the lexicon to contain a maximum of I_{tot} lexical entries, the optimal lexicon can be found by the following criterion:

$$L'_V = \operatorname{argmax}_{\{I_v, B_v^i\}} \prod_{v=1}^V \left\{ \prod_{n=1}^{I_v} \max_i p(X_v^{(n)} | B_v^{(i)}) \right\} \wedge \sum_{v=1}^V I_v \leq I_{tot} \quad (2)$$

(2) expresses that in order to create the optimal lexicon, we need to find both the optimal number of baseforms, I_v , for each keyword and the corresponding baseforms. This optimization task is a clustering problem. The expression within braces represents the clustering of training tokens given a set of baseforms. When a cluster has been defined, the optimal baseform can be found by (1). The optimization is constrained by a maximum number of baseforms in the lexicon, and optionally, a maximum number of baseforms that is allowed for any individual keyword.

The following two-stage algorithm gives a sub-optimal solution to the optimization problem[7]:

Stage 1: For each word v in the lexicon, find the best representation using $i=1, \dots, I_{max}^{(v)}$ baseforms, and the associated likelihood of the training tokens.

- 1) Assign all training tokens to a single cluster ($i=1$). Find the optimal baseform using (1). Calculate the likelihood of the training tokens by forced alignment with the optimal baseform.
- 2) Split the cluster with the lowest likelihood by selecting the two cluster members furthest apart as new cluster centers. Reassign cluster membership by using the phone strings obtained by phone recognition as the baseforms defining the new clusters. Increment i .
- 3) Find new optimal baseforms for all clusters using (1). Re-cluster the training data for the new baseforms. Calculate the likelihood of the training utterances by forced alignment to the cluster's optimal baseform. Iterate until convergence.
- 4) If $i < I_{max}^{(v)}$ repeat from 2)

DARPA RM %words correct	1 mix	2 mix	3 mix
Baseline	78.3	84.3	87.2
Optimized, $\bar{I}_v = 1.1$	81.1	85.4	88.1
Optimized, $\bar{I}_v = 1.3$	82.2	86.0	88.1

Table 1: Performance of pronunciation variation modeling vs. standard lexicon on DARPA RM task. CI models with 1-3 mixture components. \bar{I}_v is the average number of baseforms per keyword [7].

Stage 2: Determine the optimal number of baseforms for each word.

- 1) Assign a single baseform to each vocabulary word. Find the associated total likelihood of the training data from the single baseform likelihoods for each word, as calculated in Stage 1.
- 2) Identify the word where assigning an extra baseform will increase the total likelihood most. If the number of baseforms of that word is less than the maximum number of baseforms, update the lexicon by replacing the existing baseforms for that word with the new set of baseforms. Update the total likelihood.
- 3) If the total number of baseforms assigned to the lexicon is less than the predesigned maximum, repeat from 2).

In an experiment on the Resource Management task[7], it was shown that the above procedure improved the recognition rates when the acoustic models were relatively simple, i.e. context independent phone models with 1-2 mixture components. However, when the number of mixture components increased, the performance improvement was very small. However, it should be noted that the pronunciation dictionary for this task originally was manually constructed with the aim of minimizing recognition errors. Also, the task contains only read speech.

Another experiment using this method has been performed on the task of isolated name recognition. The corpus contained both native English speakers and non-native speakers. The dictionary contained both English and foreign names. For the sub-task of recognizing the first names the vocabulary contained approximately 1.800 words. The acoustic models were task independent models trained on the WSJ0 corpus, and the standard pronunciation lexicon was created using a TTS front-end. The standard pronunciation lexicon contained pronunciation variants, with an average of 2.25 baseforms/keyword. Applying the proposed procedure for

pronunciation variation modeling, the word error rate was reduced by nearly 3% absolute.

3.3 Acoustic subwords

Most automatic speech recognizers employ phonemic subword units, which are based upon an abstract linguistic description of the language. However, the analysis of the actual speech signal is acoustically based. The resulting systems are thus neither phonetically nor acoustically consistent, but is instead a hybrid of two methodologies. In order to create a consistent acoustic framework, there have been several attempts to utilize acoustically based subword units (e.g. [8], [9], [10]). The major obstacle to the use of acoustic subwords has been the lack of a linguistic interpretation of the units, making the creation of a pronunciation lexicon a difficult problem.

In [11] it was proposed to apply the maximum likelihood approach to pronunciation modeling to an acoustic subword based speech recognizer. Initially, the training data were segmented into acoustically stable segments using Constrained Clustering Segmentation. For a training utterance, this can be formulated as finding the set of segment boundaries $\{b_i\}$ that minimizes the distortion

$$D_{tot} = \sum_{j=0}^{J-1} \sum_{n=b_j}^{b_{j+1}-1} d(\mathbf{x}_n, \mathbf{c}_j) \quad (3)$$

where \mathbf{c}_j is the centroid of the j 'th segment, and $d(\mathbf{x}, \mathbf{y})$ is the distance between vectors \mathbf{x} and \mathbf{y} . After segmentation, N acoustic subword classes are defined by clustering of the segment centroids. These classes constitute the acoustic subword unit inventory. The identity of the unit is simply the cluster index, and the segments are labeled with the index of the cluster to which their centroids are assigned. Now, initial models can be estimated for each unit. Since the segments will be acoustically stable, single-state HMMs are sufficient.

A pronunciation lexicon can then be generated using the ML pronunciation modeling as described in 3.1. We can then iteratively improve the HMMs and the pronunciation lexicon. Experiments on the DARPA Resource Management task showed that this method could obtain comparable results to traditional, phone-based models, see figure 1.

3.4 Indirect data-driven pronunciation variation modeling

Widespread deployment of ASR requires that the technology is able to handle pronunciation variations. Non-native speakers constitute a group where it is difficult to obtain a good model of the pronunciations. The pronunciation will depend on the linguistic background of the speaker and on the individual's ability to adapt to the new language. The pronunciations from this group will thus be quite variable.

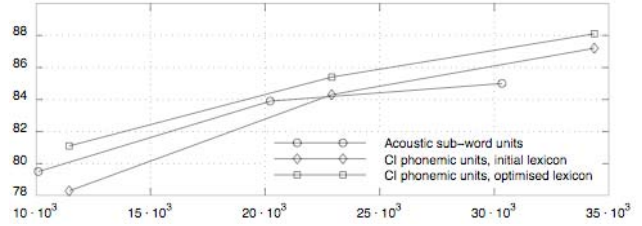


Figure 1: Word correct rates vs. total number of free parameters for acoustic modeling. Acoustic subword units vs. phonemic units with and without optimized lexicon[11].

In [12] a completely automated approach to indirect data-driven modeling of pronunciation variation modeling was proposed for this problem. A reference lexicon and corresponding utterance transcriptions were obtained using a TTS front-end. The basic steps of the procedure consisted of the following:

- 1) Alternative transcriptions were generated using a speaker independent phone recognizer trained on native speakers.
- 2) The reference and alternative transcriptions were aligned using a dynamic programming procedure using association strength[13] between phones as the basis for the cost measure for substitutions.
- 3) The mapping between the transcriptions were described with a rule set taking into account the immediate left and right phonemic context of the reference phone discarding infrequent rules.
- 4) The resulting rule set was pruned using a metric based on the acoustic log likelihood of the training data, retaining the rules that gave the highest log likelihood improvement.
- 5) A final pruning to reduce confusability was performed by restricting the rule set to allow only a single rule per source segment.

The procedure was tested on the Wall Street Journal adaptation and test set for non-native speakers of American English. The resulting rule set contained 19 rules and gave a WER of 28.3%, an improvement of the baseline performance which had a WER of 29.2%

3.5 Cross-lingual ASR

For many languages, the availability of speech data suitable for training speech recognizers is scarce. A possible way to mitigate this problem is to utilize existing language resources from other *Source* languages for porting ASR technology to the *Target* language. In order to efficiently use the source language resources, it is necessary to have a mapping between the acoustic units of the two languages.

Approaches to obtaining such a mapping has included knowledge-based methods, exploiting linguistic knowledge of the representational units of the two

languages, and data-driven methods typically utilizing confusion matrices or entropy based measures.

In [14] a method for obtaining such a mapping based on pronunciation modeling was proposed. The procedure is based on these steps:

- 1) Using acoustic models for the source language, obtain a pronunciation lexicon of the *syllables* of the target language in terms of the source language symbols. To this end, data driven pronunciation modeling (ML modeling, see section 3.1) is applied on syllables for which there are sufficiently many training examples in the training data in the target language. This will produce a lexicon with incomplete coverage.
- 2) Use a reference lexicon for the target language and the new pronunciation lexicon to derive a set of rules that map the target language phones to the source language. Prune the rules.
- 3) Apply the rule set to produce the missing lexical entries for keywords (syllables) in the target language in terms of source language symbols.
- 4) Recognition of speech in the target language can now be performed using the pronunciation lexicon and source language acoustic models.

Pilot experiments on Indonesian using English as the source language showed that this method performed better than knowledge-based and confusion matrix based methods for cross-lingual modeling. The performance of all the cross-lingual systems were however rather poor, and it is clear that more research is needed in this area.

4. CONCLUSIONS

We have presented an overview of approaches to pronunciation modeling for speech technology, giving some examples of methods. The examples have been restricted to automatic speech recognition, which reflects the research activity in this area. Applications of pronunciation modeling in TTS, apart from knowledge-based approaches, are few, but it is likely that pronunciation modeling could e.g. be an aid in the production of databases for unit selection synthesis. This paper has not attempted to give a comprehensive tutorial review. For more literature on pronunciation modeling, the interested reader is referred to [15],[16],[17] and [18].

5. ACKNOWLEDGEMENTS

The author's own work on pronunciation modeling has spanned a rather long time period. I would like to acknowledge the collaboration with a number of people: Dr. Frank K. Soong and Prof. Kuldip K. Paliwal in the "early years", Dr. Trym Holter and Dr. Ingunn Amdal later on, and most recently with Terry Martin.

11. REFERENCES

- [1] H. Strik, C. Cucchiaroni: "Modeling Pronunciation Variation for ASR: A Survey Of The Literature", *Speech Communication*, Vol. 29, No. 2-4, pp. 225-246, Nov. 1999
- [2] I.Amdal, E.Fosler-Lussier: "Pronunciation Variation Modeling in Automatic Speech Recognition", *Elektronikk*, 2.2003, pp. 70-82, 2003
- [3] A.W.Black, K.A.Lenzo: *Building Synthetic Voices*, <http://www.festvox.org/bsv/>
- [4] E.Fosler-Lussier, N.Morgan: "Effects of Speaking Rate and Word Frequency on Pronunciations in Conversational Speech", *Speech Communication*, Vol. 29, No. 2-4, pp. 137-158, Nov. 1999
- [5] M. Riley et al.: Stochastic Pronunciation Modeling from Hand-Labeled Phonetic Corpora", *Speech Communication*, Vol. 29, No. 2-4, pp. 209-224 Nov. 1999
- [6] T.Svendsen, F.K.Soong, H.Purnhagen: "Optimizing Baseforms for HMM-Based Speech Recognition", *Proc. Eurospeech-95*, pp. 783-786, Madrid, Spain, 1995
- [7] T.Holter, T.Svendsen: "Maximum Likelihood Modeling of Pronunciation Variation", *Speech Communication*, Vol. 29, No. 2-4, pp. 177-192, Nov. 1999
- [8] J.G.Wilpon, B.-H.Juang, L.R.Rabiner: "An Investigation on the Use of Acoustic Sub-Word Units for Automatic Speech Recognition", *Proc. ICASSP -87*, pp. 821-824, Dallas, 1987
- [9] C.-H.Lee, F.K.Soong, B.-H.Juang: A Segment Model Approach to Speech Recognition", *Proc. ICASSP-88*, pp.501-504, New York, 1988
- [10] T.Svendsen, K.K.Paliwal, E.Harborg, P.O.Husøy: "An Improved Sub-Word Based Speech Recognizer", *Proc. ICASSP-89*, pp. 683-686, Glasgow, 1989
- [11] T.Holter, T.Svendsen: "Combined Optimization of Baseforms and Model Parameters in Speech Recognition Based on Acoustic Subword Units", *Proc. 1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 199-206, Santa Barbara, 1997
- [12] I.Amdal, F.Korkmazskiy, A.C.Surendran: "Joint Pronunciation Modeling of Non-Native Speakers Using Data-Driven Methods", *Proc. ICSLP-2000*, pp. 622-625, Beijing, 2000
- [13] I.Amdal, F.Korkmazskiy, A.C.Surendran: "Data-Driven Pronunciation Modeling for Non-Native Speakers Using Association Strength Between Phones", *Proc. ISCA ITRW ASR2000*, pp. 85-90, Paris, 2000
- [14] T.Martin, T.Svendsen, S.Sridharan: "Cross-Lingual Pronunciation Modeling for Indonesian Speech Recognition"; *Proc. Eurospeech-03*, pp. 3125-3128, Geneva, 2003
- [15] Proc. ESCA Tutorial and Research Workshop on *Modeling Pronunciation Variation for Automatic Speech Recognition*, Rolduc, May 1998
- [16] Proc. ISCA Tutorial and Research Workshop on *Adaptation Methods for Speech Recognition*, Sophia-Antipolis, Aug. 2001
- [17] Proc. ISCA Tutorial and Research Workshop on *Pronunciation Modeling and Lexicon Adaptation for Spoken Language*, Estes Park, Sept. 2002
- [18] Special Issue on Modeling Pronunciation Variation for Automatic Speech Recognition, *Speech Communication*, Vol. 29, No. 2-4