PRONUNCIATION MODELING FOR SPONTANEOUS SPEECH RECOGNITION USING LATENT PRONUNCIATION ANALYSIS (LPA) AND PRIOR KNOWLEDGE

Che-Kuang Lin and Lin-Shan Lee

National Taiwan University, Taipei, Taiwan
kimchy@speech.ee.ntu.edu.tw , lslee@gate.sinica.edu.tw

ABSTRACT

In this paper, we propose a new framework for pronunciation modeling, in which the search algorithm tries to focus primarily on the clearly-pronounced portion of speech, while deemphasizing the observations of the slurred portion. This is based on the prior analysis that the pronunciation variation has to do with the predictability and the importance of the words in the spoken utterances, which may be estimated to some extent. We define a set of pronunciation-related features and develop a Latent Pronunciation Analysis (LPA) to estimate the "latent pronunciation states" in the speech. The LPA probabilities, pronunciation-related features and another set of prior knowledge obtained from two distance measures between phonemes are integrated in a SVM classifier to produce a "pronunciation variation indicator" for each frame, based on which the Viterbi decoding was performed. Very encouraging initial results on Mandarin spontaneous speech were obtained in preliminary experiments.

Index Terms— Pronunciation variation, spontaneous speech, speech recognition, Probabilistic Latent Semantic Analysis, Distance metrics.

1. INTRODUCTION

The pronunciation variation present in spontaneous speech is one of the major problems in spontaneous speech recognition. Various techniques have been proposed to deal with this problem. Articulatory features have been used for feature-based pronunciation modeling [1]. Feature spaces covering longer time segments were also used to implicitly model pronunciation variation due to coarticulation [2]. Another family of popular approaches employed multiple-pronunciation dictionaries which include pronunciation variants in addition to the canonical pronunciation of words to capture the pronunciation variation [3,4]. However, the added variants also introduce extra lexical confusion because of the increase in words sharing identical or similar pronunciations [3,4]. Another group of methods use "automatically derived units" to make the acoustic model optimized with respect to pronunciation variants. Ergodic hidden Markov models (EHMM) for such sub-word units were also developed as probabilistic pronunciation networks for lexical words [5]. Still other approaches tried to integrate models of acoustic units from non-native speakers [6], or adapted the acoustic models based on pronunciation variability analysis [7].

In this paper, we develop a new framework for handling the pronunciation variation for large vocabulary spontaneous speech recognition, where the pronunciation variation considered is mainly from the intra-speaker variability. This is based on the prior analysis that the pronunciation variation very often has to do with the predictability and the importance of the words in the spoken utterances, in addition to many other factors. We try to model the pronunciation by a set of pronunciation-related features, a set of "latent pronunciation states", and the prior knowledge constructed from two distance measures between phonemes estimated from a training corpus. These are all used in a Support Vector Machine (SVM) to obtain a "pronunciation variation indicator" for each frame of the speech signal. This indicator can then be applied in a modified search algorithm which de-emphasizes the frames of slurred phonemes but focuses more on those of clearly-produced phonemes. Very encouraging initial results are reported.

Below in Section 2, we present the framework of the proposed approach, while in Section 3, we describe the initial experimental results. The concluding remarks are finally made in Section 4.

2. PROPOSED APPROACH

2.1. Basic principle and the overall picture

Prior analysis indicated that speakers automatically adjust their articulatory efforts to accommodate the listener considering the predictability and the importance of the information carried by the spoken words [8]. The novelty of information for each spoken word varies, and phonemes are naturally hyper-articulated during points of emphasis and reduced at very predictable instants [9]. Other factors such as POS also lead to the varied articulatory efforts. However, despite all these pronunciation variation, the intelligibility of speech is seldom degraded for human listeners. This is why in this paper we propose to design the speech recognizer matching the above speech production process by focusing more on the clearly-pronounced frames while deemphasizing the error-prone parts of speech, instead of decoding them uniformly. The negative impact from pronunciation variation may thus be avoided while the lost information can be recovered based on the information in the more clearly-produced parts as well as other context information, such as those modeled by n-gram probabilities.

The proposed approach is shown in Figure 1. In the training phase, the first-pass recognition produces word graphs for each training utterance, from which a whole set of features can be extracted. Some of the features are used in the Latent

Pronunciation Analysis (LPA) to be presented below in section 2.3. Another set of prior knowledge can be directly constructed from the training corpora as discussed in section 2.4. All this information is used in the SVM classifier training. In the testing phase, the first-pass recognition is also performed to produce word graphs, from which the various features are extracted and the Latent Pronunciation Analysis (LPA) is performed. The SVM classification then produces a "pronunciation variation indicator" u(t) for each frame of speech, which has to do with the reliability of the pronunciation of the frame, to be used in the weighted Viterbi decoding.

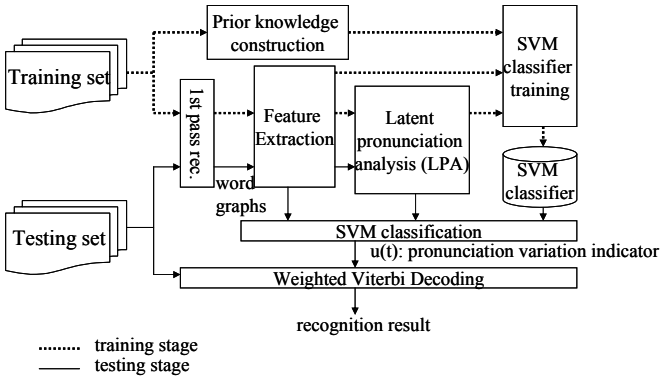

Figure 1. Overview of the proposed framework for pronunciation modeling

2.2. Feature extraction

2.2.1. Basic features extracted from word graph

The change from a baseform pronunciation into a surface form, together with the hypothesized history propagating along the search space, may lead the recognition process to generate different word hypotheses with different probability scores in the word graph. We thus try to extract a whole set of new pronunciation-related features from the word graphs produced by the first pass recognition of utterances.

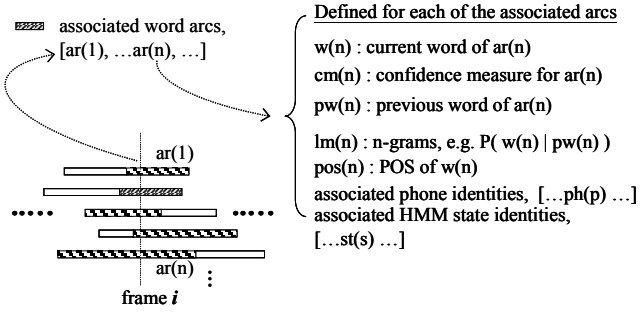

Figure 2. Basic features extracted from the word graph

For a given word graph, we can first estimate the confidence measure for each word arc in the graph using the forward-backward algorithm. Similarly applying the algorithm for the phone lattice expanded from each word arc, confidence measure for each phone can be obtained. Part-of-speech tagging is also applied throughout the word graph. As shown in Figure 2, for a frame i all the word arcs including the frame, [ar(1), **…** ar(n) **…**], are the associated word arcs, and a whole set of basic features can thus be obtained for frame i from these arcs, such as the current word w(n) for ar(n), confidence measure cm(n) for ar(n), n-gram probabilities lm(n) for ar(n), etc.

2.2.2. Word predictability features

As mentioned previously, pronunciation variation has to do with the predictability of the words. A set of features considering word predictability are thus defined for each frame. One example is the average language model score Lm over all associated word arcs, ar(n), weighted by their respective word confidence measures cm(n),

$$Lm = \sum_n cm(n)lm(n). \quad (1)$$

Another example is the average conditional entropy Ent over all associated word arcs, also weighted by the word confidence measures:

$$Ent = \sum_n cm(n)ent(n), \quad (2)$$

where ent(n) is the normalized conditional entropy for the possible word spoken at this frame, v_i ,given the previous word pw(n) of this arc, i.e.

$$ent(n) = -\{\sum_i p(v_i \mid pw(n)) \log p(v_i \mid pw(n))\} / \log |V| \quad (3)$$

where V={ v_i } is the set of all possible words spoken at the current frame, which can be defined in different ways. Another example is the averaged entropy similarly evaluated based on the confidence measures of the associated word arcs.

2.2.3. Phone level features

A whole set of phone level features can also be obtained. For example, a set of Gaussian mixtures with large enough weights from the acoustic models can be first collected. The Gaussian likelihoods for the current frame can be evaluated with no priors, and the entropy evaluated based on the posteriors can then be used to measure the confusion among acoustic units. This can be further expanded to adjacent frames on the left and right, and so on. The phone level confidence measure as mentioned in (2.2.1) can be another example. The entropy evaluated based on them may also indicate the degree of confusion.

2.3. Latent pronunciation analysis (LPA)

There can be many unknown factors behind the phenomena of pronunciation variation, referred to as "latent pronunciation states" here in this paper, for example the slurred phones for less important function words. But these latent states may be important in modeling the pronunciation. We thus propose here to analyze the "latent pronunciation states" from the observable word graph structure using the Probabilistic Latent Semantic Analysis (PLSA) framework useful in the area of information retrieval [10]. In PLSA, instead of directly counting the co-occurrence statistics between the document set $\{d_i\}$ and the term set $\{t_k\}$, a set of latent topics $\{z_l\}$ is created and the relationships between each document d_i and each term t_k are modeled by a probabilistic framework via these latent topics $\{z_l\}$:

$$P(t_k \mid d_i) = \sum_l P(t_k \mid z_l)P(z_l \mid d_i), \forall i,k, \quad (4)$$

where the probabilities were trained with an EM algorithms by maximizing the total likelihood function:

$$L_T = \sum_i \sum_k n(t_k, d_i) \log P(t_k \mid d_i), \quad (5)$$

and $n(t_k, d_i)$ denotes the frequency count of t_k in d_i.

In the Latent Pronunciation Analysis (LPA) developed here, we treat the set of word arcs associated with each frame i, [ar(1), …ar(n), …] as shown in Figure 2, as the "pronunciation documents", d_i , and each individual word arc, ar(n), including the identities of the previous word, the current word, the current phone and the HMM state for the arc at the frame as the "pronunciation term" t_k in the document d_i , while the latent topic z_l corresponds to the latent pronunciation state, as illustrated in Figure 3. All the probabilities in equation (4) can be trained with EM algorithm as in PLSA, with the frequency count $n(t_k, d_i)$ in equation (5) taken as the confidence measure for phone identity ph(p) of t_k in d_i . This is the LPA proposed here in this paper. With such a model, the probabilities of each frame d_i belonging to a pronunciation state z_l, $p(z_l \mid d_i)$, can be taken as extra features to be used below.

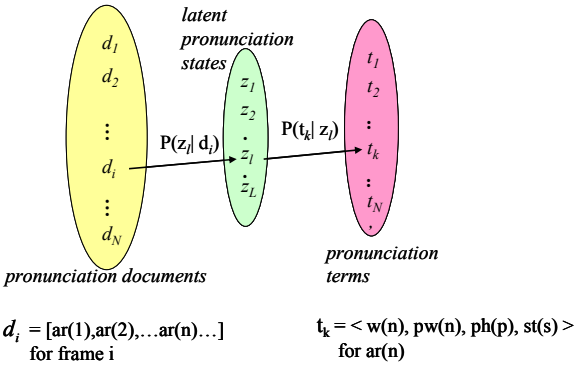

Figure 3. Latent Pronunciation Analysis (LPA)

2.4. Prior knowledge construction

A phoneme p may be recognized as another phoneme q because they are acoustically similar to each other, or because the speaker actually produced p as q due to the pronunciation variation. The phenomena may be even more complicated since the two situations may happen simultaneously. Both of these situations can be analyzed in advance from a training corpus, and such prior knowledge can be useful in pronunciation variation modeling. Here this is done based on two statistical metrics, acoustic distance and phonemic distance between each pair of phonemes [4].

The acoustic distance $d_{ac}(r;q)$ measures how likely it is that a phoneme q may be recognized as another phoneme r, and can be evaluated by, for example, the Kullback Leibler distance between two HMMs. This distance gives the possible degree of confusion between two phonemes. On the other hand, the phonemic distance describes how likely it is that a phoneme p (base form) is realized as another phoneme q (surface form),

$$d_{ph}(q;p) = -\ln[\, Pr(\, T_s = q \mid T_c = p\,)\,] \quad (6)$$

where T_s and T_c are aligned surface and canonical transcriptions respectively. Both of these two distances are asymmetric and can be used to construct confusion matrices to be used here [4]. For example, for a recognized (or produced) phoneme r (or q), the distances $d_{ac}(r;q)$ and $d_{ph}(q;p)$ for all phonemes $q \neq r$ (or $p \neq q$) can be used to evaluate entropy measures to be used in the approach here.

In addition, since the phonemic distance describes the relationship from the target phoneme p to the produced phoneme q, while the acoustic distance describes the relationship from the produced phoneme q to the recognized phoneme r, the two phases of speech production and recognition can be considered as cascaded as in Figure 4. For a speech frame recognized as a phoneme r, it may be actually produced as different phonemes q's, which in turn may be the surface forms of some other target phonemes p's. Each path in Figure 4 can thus be assigned a probability based on d_{ph} and d_{ac}. For a given phoneme r the probabilities for all paths leading to r with p=q and $p \neq q$ can be used to measure the probability of pronunciation variation given r.

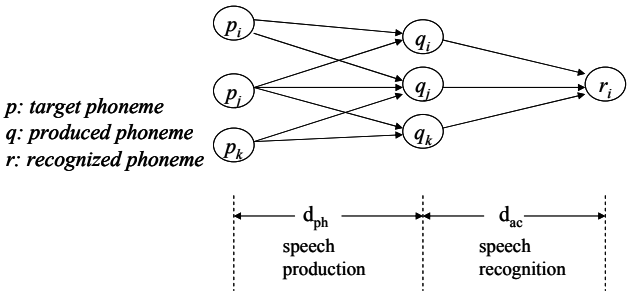

Figure 4. The cascade of speech production and recognition processes for the two distance measures.

2.5. SVM classifier training

All the features, LPA probabilities and prior knowledge presented above can be used to train the support vector machine (SVM) with a radial basis kernel. For each frame in the training set, whether the speaker actually produced the target baseform or the varied surface form (a binary decision) should be determined first. A previously developed automatic surface form generation procedure [4] was used here, in which phone-level forced recognition was performed on the training data based on the phone level confusion table obtained in the previous stage of free-phone recognition, phone alignment and error pruning. The SVM classifier training was then performed with the LIBSVM toolkit [11]. The SVM classifier then produced a "pronunciation variation indicator" u(t) (a real number between 0 and 1) for each test frame at time t.

2.6. Weighted Viterbi decoding

In the Viterbi decoding in testing, a modified observation probability, $b_j'(o_t)$, for the feature vector o_t at time t in HMM state j can be used,

$$b_j'(o_t) = b_j(o_t)^{u(t)} \quad (7)$$

where $b_j(o_t)$ is the original observation probability for the HMM. Here u(t) is between 0 and 1, because parts of the surface form may still carry some helpful information and should not be completely ignored.

3. EXPERIMENTAL RESULTS

The corpus used in this research was taken from the Mandarin Conversational Dialogue Corpus (MCDC) [12], collected from 2000 to 2001 by the Institute of Linguistics of Academia Sinica in Taipei, Taiwan. 8 dialogues out of the 30, with a total length of 8 hrs, produced by nine female and seven male speakers, were used in this research. 7.1 hours of the corpus were used in training, and the rest for testing. The test set was chosen to cover all the speakers.

Experiments (feature sets)	recall	precision	Char acc.
baseline	N/A	N/A	47.40
word predictability features	65.1	69.3	47.69
plus POS	65.3	69.1	47.75
plus phone level features	66.2	70.1	47.93
plus LPA(32states)	67.1	71.6	48.47
plus prior knowledge	69.2	72.3	48.77
"oracle"	100.0	100.0	58.99

Table 1. The performance of pronunciation variation classification and the character accuracy obtained in the speech recognition experiments.

3.1. Pronunciation variation classification

We first evaluate the performance of the pronunciation variation classification performed by the SVM classifier, which classifies each testing frame as either in baseform ($u(t)>0.5$) or in surface form ($u(t) < 0.5$). The results of recall and precision rates in the first two columns of Table 1 are based on the surface forms obtained by the previously developed surface form generation procedure mentioned in section 2.5 based on known transcriptions. The word predictability features alone produced recall and precision rates of 65.1 and 69.3 respectively. Adding the POS tags, phone level features, LPA probabilities and the prior knowledge improve the performance step by step, up to 69.2 and 72.3.

Further analysis was performed regarding the relative importance of the different features used (only the word predictability, POS, and phone level features were considered), and the five most important features were identified and listed in Table 2.

3.2. Large vocabulary spontaneous speech recognition

The baseline recognition system uses a canonical lexicon of 50K entries, a trigram language model, and an intra-syllable right context dependent Initial/Final acoustic model set (a Mandarin syllable is decomposed into two parts: Initial and Final). An "oracle" experiment is also performed, in which the surface forms obtained based on known transcriptions, as is used in the recall/precision evaluation, is used in the weighted Viterbi decoding with hard valued weight (1 for baseform and 0.2 for surface forms). Table 1 shows that in the last row of the last column the character accuracy on the test set improves from 47.40% for the baseline to 58.99% for the "oracle" case. This certainly verifies the high potential of the concept proposed here, i.e., relying primarily on the clearly-produced parts but deemphasizing the slurred parts of speech is very helpful. The actual recognition accuracy in the last column of Table 1 also improves step by step when more information is used, although the improvements achieved here seem to be relatively limited in the initial experiments, probably due to the relatively low recall/precision in Table 1. The distance from the "oracle" result may indicate the results here are still quite preliminary.

(1)The bi-gram probability of the word arc with the highest confidence measure
(2)The largest confidence measure of the word arc associated with the current frame
(3)The bi-gram probability of the word arc with the second highest confidence measure
(4)The uncertainty of the next word
(5)The entropy of the confidence measure associated with the current frame

Table 2. List of the five most important features among the set described in section 2.2.

4. CONCLUSION

A new framework for pronunciation modeling is presented. The "latent pronunciation state", prior knowledge and a whole set of features were used to calculate a "pronunciation variation indicator" for each frame to be used in weighted Viterbi decoding. Significant improvement in recognition accuracy can be achieved if perfect information about the state of pronunciation variation is available, while moderate improvements in actual recognition accuracy were obtained in initial experiments.

5. REFERENCES

[1] K. Livescu and J. Glass, "Feature-based pronunciation modeling with trainable asynchrony probabilities," *in Proc. ICSLP*, 2004.

[2] S. Dupont, C. Ris, L. Couvreur, and J.-M. Boite, "A study of implicit and explicit modeling of coarticulation and pronunciation variation," *in Proc. INTERSPEECH*, 2005.

[3] H. Strik and C. Cucchiarini, "Modeling pronunciation variation for ASR: a survey of the literature," *Speech Communication*, pp. 225-246, Nov. 1999.

[4] M.-Y. Tsai, F.-C. Chou, and L.-S. Lee, "Pronunciation modeling with reduced confusion for Mandarine Chinese using a three-stage framework," to be published in *IEEE Trans. on Speech and Audio Processing*, 2006.

[5] V. Ramasubramanian, P. Srinivas, T.V. Sreenivas, "Stochastic pronunciation modeling by Ergodic-HMM of acoustic sub-word units," *in Proc. INTERSPEECH*, 2005.

[6] G. Bouselmi, D. Fohr, I. Illina, J.P. Haton, "Fully automated non-native speech recognition using confusion-based acoustic model integration and graphemic constraints," *in Proc. ICASSP*, 2006.

[7] Y.R. Oh, J.S. Yoon, and H.K. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition," *in Proc. ICASSP*, 2006.

[8] R. Bates and M. Ostendorf, "Modeling pronunciation variation in conversational speech using prosody," *in Proc. PMLA*, 2002.

[9] B. Lindblom, *Speech Production and Speech Modeling*, chapter Explaining Phonetic Variation: A Sketch of the H&H Theory, pages 403-439. Kluwer Academic Puflishers, 1990.

[10] T. Hofmann, "Probabilistic latent semantic analysis," *in Proc. Uncertainty in Artificial Intelligence*, 1999.

[11] C.-C. Chang and C.-J. Lin, LIBSVM: A library for support vector machines, 2001. www.csie.ntu.edu.tw/~cjlin/libsvm

[12] S.-C. Tseng, "Processing spoken Mandarin corpora," *Traitement automatique des langues.* Special Issue: Spoken Corpus Processing. 45(2): 89-108, 2004.