PAPER

Sequence-Based Pronunciation Variation Modeling for Spontaneous ASR Using a Noisy Channel Approach

Hansjörg HOFMANN[†*a)], **Sakriani SAKTI**[†**], *Nonmembers*, **Chiori HORI**[†], *Member*, **Hideki KASHIOKA**[†], *Nonmember*, **Satoshi NAKAMURA**[†**], *Member*, *and* **Wolfgang MINKER**[††], *Nonmember*

SUMMARY The performance of English automatic speech recognition systems decreases when recognizing spontaneous speech mainly due to multiple pronunciation variants in the utterances. Previous approaches address this problem by modeling the alteration of the pronunciation on a phoneme to phoneme level. However, the phonetic transformation effects induced by the pronunciation of the whole sentence have not yet been considered. In this article, the sequence-based pronunciation variation is modeled using a noisy channel approach where the spontaneous phoneme sequence is considered as a "noisy" string and the goal is to recover the "clean" string of the word sequence. Hereby, the whole word sequence and its effect on the alternation of the phonemes will be taken into consideration. Moreover, the system not only learns the phoneme transformation but also the mapping from the phoneme to the word directly. In this study, first the phonemes will be recognized with the present recognition system and afterwards the pronunciation variation model based on the noisy channel approach will map from the phoneme to the word level. Two well-known natural language processing approaches are adopted and derived from the noisy channel model theory: Joint-sequence models and statistical machine translation. Both of them are applied and various experiments are conducted using microphone and telephone of spontaneous speech.

key words: *spontaneous speech, noisy channel approach, joint-sequence models, statistical machine translation*

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems perform satisfactory in closed environments. However, under more relaxed constraints when people speak freely, the recognition rates decrease [1]. In natural conversations people pronounce differently, tend to combine or even miss words out. Discourse particles (e.g. "like") or hesitation sounds (e.g. "ahm") are used to structure the sentence and have no semantic meaning. However, Riley et al. [2] consider multiple pronunciation variants as being one of the main problems of spontaneous ASR.

Several approaches have been made to resolve the multiple pronunciation problem. One attempt is to extend the dictionary manually with further pronunciation variants or to improve it by applying rule-based algorithms [3]. Nevertheless, both approaches are very time consuming and the latter one needs a significant amount of expert knowledge. Alternatively, data driven approaches may be used to model the alteration of the pronunciation on a phoneme-to-phoneme level. Decision-tree-based approaches applied by Bates et al. [4] have improved the ASR performance. Chen et al. [5] examine the effect of prosody on pronunciation and propose to use artificial neural networks (ANN) to model pronunciation variation. Livescu et al. [6] propose a feature-based pronunciation model based on a dynamic Bayesian Network (BN). Sakti et al. [7] also uses a BN technique to model the variation of the base form and the surface form of the phoneme. After applying the Bayesian Network a small performance improvement of the ASR was gained. Since the realization of the current phone does not only depend on neighboring phones the observation window should be extended. Fosler-Lussier [8] proposes to take syllabification into consideration and investigate decision tree models based on syllables. However, the word error rate increases slightly. This may be because the phonetic transformation effects induced by the pronunciation of the whole sentence are not considered yet.

In this paper, we model the sequence-based pronunciation variation using a noisy channel approach where the spontaneous phoneme sequence is considered as a "noisy" string and the goal is to recover the "clean" string of the word sequence. Two well-known natural language processing (NLP) approaches are adopted and derived from the noisy channel model to map from the phoneme to the word level: Joint-sequence models and statistical machine translation. By applying those approaches the whole word sequence and its effect on the alternation of the phonemes will be taken into consideration. Moreover, the system not only learns the phoneme transformation but also the mapping from the phoneme to the word directly. In this study, first the present ASR system recognizes the phonemes and afterwards the phonemes are mapped onto the word level with the two proposed pronunciation variation models.

In the next section the noisy channel model and its derived attempts, the joint-sequence models and the statistical machine translation, are explained. Section 3 describes the conducted experiments where the approaches have been applied on various spontaneous speech corpora (clean and telephone speech data) and evaluated. Afterwards, the experimental results are presented, both approaches compared and finally conclusions are drawn.

Manuscript received November 16, 2011.

[†] The authors are with the National Institute of Information and Communications Technology, Kyoto-fu, 619–0289 Japan

[††] The author is with the Department of Information Technology, University of Ulm, Germany

[*] Presently, with the Department of Information Technology, University of Ulm, Germany.

[**] Presently, with Nara Institute of Science and Technology, Nara, Japan

a) E-mail: hansjoerg.hofmann@uni-ulm.de

DOI: 10.1587/transinf.E95.D.2084

2. The Noisy Channel Approach

2.1 The Noisy Channel Model for ASR

As illustrated in Fig. 1 spoken language processing may be considered as a noisy channel problem. Here, a word sequence W considered as "clean" input is transmitted via a noisy channel where it gets modified and transformed onto a different domain, the speech signal $\underline{X}$.

The goal of the ASR is to decode the speech signal output of the noisy channel to recover the most likely input word sequence $\hat{W}$. Mathematically, this task can be formulated as follows:

$$\hat{W} = \underset{W \in \Omega}{\operatorname{argmax}}\ P(W|\underline{X}). \tag{1}$$

In this Equation, Ω denotes the set of word sequences.

Under constraint conditions, i.e., read speech with high-quality microphone recordings, the ASR task to recover the most likely input word sequence $\hat{W}$ can be done straightforwardly. However, the "noisy" speech variability increases significantly in case of spontaneous speech. Consequently, the ASR may not be able to recover the "clean" word string W anymore. Therefore, additional efforts have to be made to support the ASR. Those efforts will be described in the following section.

2.2 The Proposed Noisy Channel Model for Spontaneous ASR

In this study, we propose to split the task $P(W|\underline{X})$ into two subtasks as follows (see Fig. 2):

The first task is used to handle the acoustic variability of the spontaneous speech signal which is achieved by the ASR. Here, the ASR task is only to recover the most probable phoneme sequence $\hat{F}$ given the "noisy" spontaneous speech signal $\underline{X}$ according to Eq. (2):

$$\hat{F} = \underset{F \in \Phi}{\operatorname{argmax}}\ P(F|\underline{X}) \tag{2}$$

where Φ denotes the set of phoneme sequences F.

The second task is to deal with the present pronunciation variability. In this case, the task is to recover the most probable word sequence $\hat{W}$, given the phoneme sequence $\hat{F}$ from the ASR output, according to Eq. (2):

$$\hat{W} = \underset{W \in \Omega}{\operatorname{argmax}}\ P(W|\hat{F}). \tag{3}$$

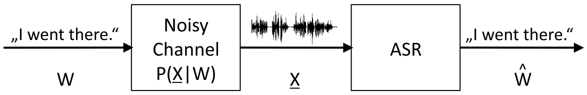


Fig. 1 The Noisy Channel Model for ASR.

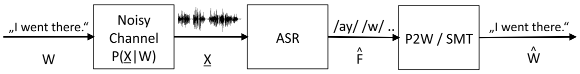


Fig. 2 Applying the noisy channel approach on spontaneous ASR.

In order to solve the second task on handling the pronunciation variation, we attempt to adopt two well-known NLP approaches which will be explained in the next two subsections.

2.2.1 Joint-Sequence Model

By applying Bayes' rule, Eq. (3) can be rewritten to[†]:

$$\hat{W} = \underset{W \in \Omega}{\operatorname{argmax}}\ P(W|F) = \underset{W \in \Omega}{\operatorname{argmax}}\ \frac{P(W,F)}{P(F)} = \underset{W \in \Omega}{\operatorname{argmax}}\ P(W,F). \tag{4}$$

We adopted the idea of Bisani et al. [9] who tried to find the most likely pronunciation of a certain word given its written form. In their report, they search for the most likely pronunciation F given its orthographic form G. This task can also be formalized using Bayes' rule:

$$\hat{F} = \underset{F \in \Phi}{\operatorname{argmax}}\ P(F|G) = \underset{F \in \Phi}{\operatorname{argmax}}\ \frac{P(F,G)}{P(G)} = \underset{F \in \Phi}{\operatorname{argmax}}\ P(F,G) \tag{5}$$

where Φ denotes the set of phoneme sequences and $\hat{F}$ denotes the most likely phoneme sequence. Bisani et al. use joint-sequence N-gram models to achieve the mapping from the orthographic to the phoneme level which is known as G2P.

By reversing Bisani et al's idea and using phoneme sequences as a source language and word sequences as a target language, the joint-sequence N-gram models can be used to compute the most probable word sequence $W \in \Omega$ given an input phoneme sequence F (as formalized in Eq. (4)). From now on, we call this approach P2W.

The joint-sequence N-gram models are trained with parallel matching pairs of text data from the input and the output language. A phoneme-word joint multigram is a pair $Q = (F, W) \in \Upsilon \subseteq \Phi \times \Omega$ of a phoneme sequence and a word sequence of possibly different lengths. The terms q_k, f_k and w_k are used to described the k-th component of $Q = (F, W)$. For example, a possible sequence Q of phoneme-word pairs for the short utterance "we will go to get her" is illustrated in Fig. 3.

As can be seen in the example, the words and the

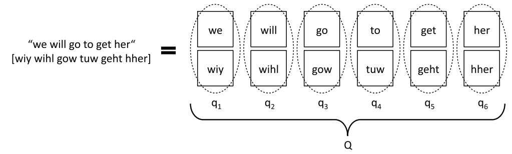


Fig. 3 Co-segmentation 1 of the example utterance.

[†] In Eq. (3) $\hat{F}$ is used instead of F. The following sections use F as variable denoting the phoneme sequence output of the ASR.

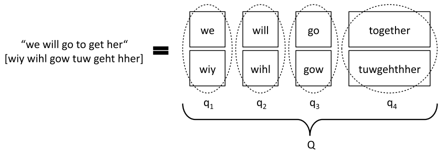


Fig. 4 Co-segmentation 2 of the example utterance.

phoneme sequences are grouped in an equal number of segments which is called $co-segmentation$. For a given input and output string there might be different ways to segment the entities. Therefore the segmentations may not be unique. A different segmentation of the above example is shown in Fig. 4.

Because of this ambiguity the joint probability $P(W, F)$ of Eq. (6) is computed by summing over all probability values of the matching sequences Q:

$$P(W, F) = \sum_{Q \in C(F,W)} P(Q). \tag{6}$$

$Q \in \Upsilon$ is a sequence of phoneme-word pairs and $C(F, W)$ is the set of all existing co-segmentations of F and W:

$$C(F, W) := \left\{ Q \in \Upsilon | \begin{array}{c} f_{q_1} \smile f_{q_2} \smile \ldots \smile f_{q_V} \\ w_{q_1} \smile w_{q_2} \smile \ldots \smile w_{q_V} \end{array} \right\} \tag{7}$$

where $V = |Q|$ denotes the length of the phoneme-word pair sequence and $\smile$ symbolizes the concatenation of the single entities. The probability distribution $P(W, F)$ can now be computed with the probability distribution $P(Q)$ over several phoneme-word pairs sequences Q. Those sequences $Q = q_1, \ldots, q_V$ can be modeled with a standard N-gram approximation:

$$P(q_1^V) \simeq \prod_{i=1}^{V+1} P(q_i | q_{i-1}, \ldots, q_{i-N+1}). \tag{8}$$

To model special phenomena at the beginning and at the end of an utterance, positions $i < 1$ and $i > V$ are also taken into consideration. The segmentation algorithms and model estimation can be adopted from the original grapheme-to-phoneme approach without any modification. Further details about these algorithms can be found in [9].

2.2.2 Machine Translation

To derive the second approach, again, Eq. (3) has to be simplified to:

$$\hat{W} = \underset{W \in \Omega}{\text{argmax}}\ P(W|F) = \underset{W \in \Omega}{\text{argmax}}\ P(F|W) \cdot P(W) \tag{9}$$

A similar formulation has been applied by Koehn et al. [10] who define a phrase translation model based on the noisy channel approach. The authors translate a foreign sentence f into an English sentence e using the following formula:

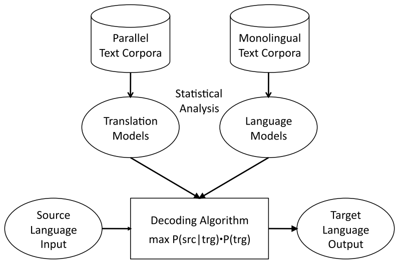


Fig. 5 Basics of SMT framework.

$$\underset{e}{\text{argmax}}\ P(e|f) = \underset{e}{\text{argmax}}\ P(f|e) \cdot P(e). \tag{10}$$

In their report, the right hand side of Eq. (10) can be represented as a translation model $P(f|e)$ and a language model (LM) $P(e)$.

In our work, this idea is adopted and instead of translating one language into another, statistical machine translation (SMT) is used to translate the recognized phonemes into words. According to the very right hand side of Eq. (9), the SMT has to compute the most probable word sequence W given an input phoneme sequence F. In this case, $P(W)$ represents the probability of the word sequence W provided by the SMT LM of the target language. $P(F|W)$ denotes the likelihood of the phoneme sequence F given the word sequence W, represents the transition from the phonemic to the word representation and is computed by the translation model. The framework of the SMT system is shown in Fig. 5.

The SMT system is trained with parallel matching pairs of text data from the input and the output language. While testing the translation system the SMT evaluates each proposed hypothesis by assigning a score according to the statistical model probabilities. During the translation process all possible hypotheses are considered and finally the path with the highest score is chosen as a result.

3. Experimental Setup

3.1 Applied Tools in the Experiments

3.1.1 ASR

NICT's ASR system uses the following features and algorithms for training and testing: A frame length of a 20 ms Hamming window, a frame shift of 10 ms and 25 dimensional feature parameters consisting of 12-order MFCC, delta MFCC and log power, are used. For building the acoustic model (AM), at the beginning each phoneme consists of a 3-state HMM. By applying a successive state splitting (SSS) algorithm based on the minimum description length (MDL) the optimum state level HMnet is obtained. Further information about the MDL-SSS algorithm may be obtained from [11].

Table 2 Buckeye corpus data composition of this experiment.

Speech Type	Data Set	Hours	Words	Usage
Spontaneous	Buckeye Train. Set	36	8741	Spont. Speech AM & LM Train, Read Speech AM Adapt
	Buckeye Train. Set	7-10	50-250	SMT & P2W Train
	Buckeye Test Set	0.2	50-250	Spont. & Read Speech ASR Test, P2W & SMT Test

Table 1 Wall Street Journal data composition of this experiment.

Speech Type	Data Set	Hours	Words	Usage
Read	WSJ Train. Set	60	4989	Read Speech ASR Train
	WSJ Test Set	0.2	4989	Read Speech ASR Test

3.1.2 P2W

The phoneme-to-word conversion system is based on Sequitur G2P[†] open source tools which is restricted to 255 words. While training, the order of the N-gram LM can be increased by ramping up the system iteratively.

3.1.3 SMT

The SMT system applies phrase-based SMT techniques [10]. It utilizes a 3-gram LM which can be tuned up onto a 5-gram LM. The statistical models of our system were trained with special toolkits for language modeling [12] and word alignment [13]. The translation process is performed by a tool called CleopATRa [14].

3.2 Data Corpora

Three speech data corpora have been used for our research. The Wall Street Journal (WSJ) corpus consists of recordings of read speech data. The Buckeye corpus contains spontaneous speech data recorded using high-quality microphones and the Switchboard corpus is based on spontaneous speech recorded from telephone conversations.

3.2.1 Wall Street Journal

The WSJ corpus (WSJ0 and WSJ1) [15] provides speech data by different English speakers who had to read newspaper texts paragraphs. The recordings were conducted in prepared rooms and the speakers wore headset microphones, therefore, the sample rate is 16 kHz.

The training set consists of 60 hours of speech and the so-called WSJ test consists of 215 utterances [16]. The WSJ test is a 5k Hub test set.

The WSJ corpus is used for two purposes in our work. First, it serves as baseline to show how well the ASR performs when recognizing speech recorded under constrained conditions such as read speech. Second, it is used to build a base AM which will be later adapted to and retrained on spontaneous speech. The data composition and its usage can be obtained from Table 1.

3.2.2 Buckeye

One of the spontaneous speech corpora is the Buckeye Corpus [17]. It is one of the richest sources of clean speech data including pronunciation transcriptions in conversational speech that is available in English. This corpus is composed of 40 native American English speakers from Ohio who had free conversations and expressed their opinions about everyday topics such as politics, sports, traffic, schools. The work of Fasold [18] suggests that such a sample is large enough to cover the interspeaker variability of the speech community. The 16 kHz recorded conversations have been transcribed and phonetically labelled. The corpus contains approximately 300k words and 9600 different words in total.

The original audio files consist of entire conversations of the speakers. Therefore, in this study, for convenience, we segmented the audio files at reasonable positions into phrases or segments in record length. As a result approximately 40k utterances (roughly 40 hours of speech) were obtained which are divided into 36 speakers (36390 utterances) used for training and 4 speakers (3385 utterances) used for testing.

Due to the word restriction of the used P2W engine, the data set has been segmented statistically into three different small vocabulary tasks: 50, 100 and 250 words[††]. In each case, 200 utterances were randomly selected from the data set partition and were used for evaluation.

As shown in Table 2 the Buckeye corpus is used for several purposes in our work. The whole training set has been employed to built an ASR (AM and LM) based on spontaneous speech. Furthermore, it is utilized to adapt and to retrain the read speech AM to spontaneous speech. The small vocabulary training set selection has been used to train the P2W engine. To compare the results properly the SMT has also been trained on the small vocabulary training set.

With the 200 test set utterances, the new ASR systems and the proposed approach have been tested on spontaneous speech data.

3.2.3 Switchboard

The second spontaneous speech data set used for our work,

[†] http://www-i6.informatik.rwth-aachen.de/web/Software/g2p.html

[††] Although the SMT engine can handle a large vocabulary range the segmentations are also used for testing the SMT since comparable results are desired.

Table 3 Switchboard corpus data composition of this experiment.

Speech Type	Data Set	Hours	Words	Usage
Spontaneous	Hand Transcribed Switchboard	4	3843	Spont. Speech AM & LM Train, Read Speech AM Adapt
	SVitchboard1 C&D&E	4-7	50-250	Spont. Speech LM Train, Read Speech AM Adapt P2W & SMT Train
	SVitchboard1 A&B	0.2	50-250	Spont. & Read Speech ASR Test, P2W & SMT Test

was obtained from two subsets of the Switchboard corpus [19]. It consists of spontaneous telephone conversations (sample rate 8 kHz) and has a significant amount of pronunciation variability [7]. The first subset has been phonetically hand-transcribed by ICSI[†] and consists of 4 hours (5117 utterances) of data. The talkers spoke approximately 4k words in total. The second one is SVitchboard1 (The Small Vocabulary Switchboard Database) [20] which consists of statistically selected utterances from the whole Switchboard corpus. SVitchboard1 only contains word transcriptions of the spoken utterances. This data set has already been segmented in small vocabulary tasks from 10 words up to 500 words. Each segmentation has been further partitioned into 5 subsets (A-E).

Table 3 shows the data composition and its usage of the Switchboard data corpus in our work. Due to the P2W word restriction we only use the three subsets 50, 100 and 250 of SVitchboard1. To adapt the read speech base AM the ICSI subset and the SVitchboard1 partitions C, D and E have been used. The new spontaneous ASR system has also been built based on the ICSI and the SVitchboard1 subset. The new spontaneous LM has been created based on both data sets. However, since SVitchboard1 does not provide phonetical transcriptions only the ICSI subset has been used for training the spontaneous AM. The SMT and the P2W system have only been trained on partitions C, D and E of Svitchboard1 because of the vocabulary restriction of P2W.

The partitions A and B of SVitchboard1 have been employed to test the different ASR systems and the proposed approach on spontaneous speech. 200 utterances from each word range subset were selected for testing.

3.3 Model Training

3.3.1 ASR Training

As described in Sect. 2.2, the ASR task is to recover the most likely phoneme sequence $\hat{F}$ given the speech signal. However, as the phoneme recognition accuracy of the ASR system is unsatisfactory, in practice, the phoneme string $\hat{F}$ is obtained from the recognized word sequence of the ASR. Two different types of AMs were built:

(a) Baseline AM
Read speech AMs based on the WSJ corpus, described in Sect. 3.2.1, were built to obtain a baseline model. The WSJ corpus contains 16 kHz sampled speech data. An 8 kHz WSJ data set has been created by down sampling the data to 8 kHz. Based on those data sets (see Table 1), two baseline AM sets were trained, consisting of a total number of approximately 2000 states (denoted as "WSJ").

(b) Adaptation
Each of the two WSJ based AMs were adapted to the conversational speech data of the Buckeye and the Switchboard corpus using the data described in Table 2 and 3 by applying the standard maximum a posteriori (MAP) adaptation (denoted as "WSJadaptBuck" and "WSJadaptSWB").

(c) Spontaneous AMs
Two AMs based on spontaneous speech data (see Tables 2, 3) were trained, consisting of less than approximately 1000 states. (denoted as "Buckeye" and "Switchboard").

For each of the 6 AMs four variants with different Gaussian mixture numbers (5, 10, 15, 20) were built. In total, 32 AMs were created during training.

The LM of the read speech ASR is based on the read speech training data. When tests are conducted which include AMs built on the Buckeye corpus, a LM based on the data of Table 2 is used. When an ASR is tested which uses an AM based on the Switchboard data, the applied LM is trained on the data of Table 3.

3.3.2 P2W Training

The P2W system was trained on the spontaneous text data (see Table 2 and 3) with the phoneme as a source and the word as the target. Here, dictionary based canonical phoneme sequences and hand-labelled surface phoneme sequences were used (see Tables 2 and 3). While increasing the order of the P2W N-gram LM, better results could be achieved. The best results were obtained when using a 7-gram or 8-gram model. Further incrementation of the order led to a saturation level.

3.3.3 SMT Training

The SMT was trained on the spontaneous text data with the phoneme as a source and the word as the target. Again,

[†] International Computer Science Institute, http://www.icsi.berkeley.edu/

Table 4 Recognition accuracy of different acoustic models in %.

AM	Testset	Mixtures			
		5	10	15	20
WSJ	WSJ	88.7	89.1	89.2	90.8
WSJ	Buckeye	29.5	29.6	33.4	29.6
WSJadaptBuck	Buckeye	38.3	39.7	43.3	41.5
Buckeye	Buckeye	55.6	55.6	57.9	54.5
WSJ	Switchboard	35.8	38.6	36.9	40.1
WSJadaptSWB	Switchboard	51.2	51.3	49.7	50.1
Switchboard	Switchboard	55.2	58.3	58.7	58.1

dictionary based canonical phoneme sequences and hand-labelled surface phoneme sequences are used. The used data sets are illustrated in Tables 2 and 3. The SMT engine only allows a trigram or a 5-gram model. By testing, it was revealed that the 5-gram model achieved better results and was used for the further experiments.

3.4 Results

3.4.1 Baseline ASR System

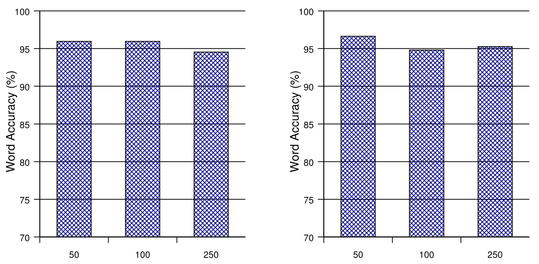
To determine the best baseline the new AMs have been tested on spontaneous speech. As reference the WSJ AM has also been tested on the WSJ test set. The word accuracy (WA) results of the different AMs tested on read speech and the Buckeye 50 words test set are shown in Table 4.

According to Table 4 the adaption helped to improve the read speech AM. The AMs which have been trained only on spontaneous speech data ("Buckeye" and "Switchboard") achieved the best results. In average across all variants, the Buckeye AM raised the WA from 30.6% to 55.9% and achieved a relative improvement of 83.2%. The Switchboard AM enhanced the WA from 37.8% to 57.6% which accords to a relative improvement of 52.4%.

The results showed the same characteristics for both corpora. Presenting all results would go beyond the scope of this paper, therefore, only a representative sample of the results will be illustrated.

3.4.2 Performance of the Proposed Approach

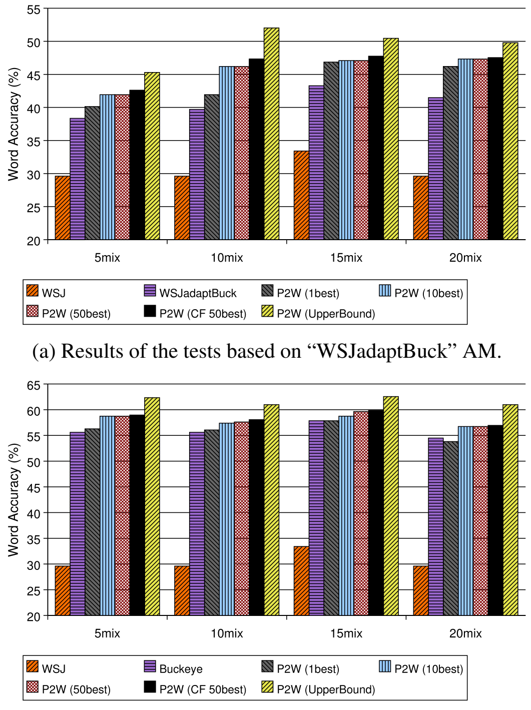
Before P2W and SMT can be applied on the ASR output the performance on correct phoneme transcriptions has to be determined. Here, the input utterances were dictionary based canonical phoneme sequences obtained from the test utterances. They have to be mapped on the word level by the proposed approach. Although while training, we mixed canonical and surface form phoneme sequences, the P2W and SMT performs well on correct phoneme sequences of the test data. The spontaneous phoneme sequences only cause little confusion and the P2W still reaches up to 96% accuracy and the SMT up to 97%. By increasing the word range from 50 to 250 words the performance decreases only slightly and seems to reach a saturation level (see Fig. 6).



(a) P2W trained and tested on Buckeye.

(b) SMT trained and tested on Buckeye.

Fig. 6 Results of testing the P2W on correct phoneme transcriptions of the different test sets.



(a) Results of the tests based on "WSJadaptBuck" AM.

(b) Results of the tests based on "Buckeye" AM.

Fig. 7 P2W test results based on the 50words Buckeye test set.

3.4.3 ASR Improvement with the Proposed Approach

In this evaluation setup, P2W and SMT are applied after the ASR is conducted. The ASR outputs the most likely words sequence and also the according phoneme sequence. However, for further processing only the phoneme strings are used. First, the results when applying P2W are presented, followed by the results of SMT.

(a) ASR Improvement Using P2W

For the 50 words test set the results of all mixture variants (based on the "WSJadaptBuck" and the "Buckeye" AM) are illustrated in Fig. 7. Given the first best path of the ASR output (P2W (1best)) the system achieves to improve all mixture variants of "WSJadaptBuck". Concerning "Buckeye", the 5mix and the 10mix variant could be improved.

However, here is only the best result of the speech recognition considered and mapped onto the word level. For further improvement we keep the whole lattice of the ASR result. Here, we apply the P2W on the n-best (10best and 50best) list generated from the ASR which results in unique word sequences (P2W (10best) and P2W (50best)). Now, all mixture variants of both AMs could be improved. In the 50words test set, the 10mix variant of "WSJadaptBuck" achieves with 16.4% the highest relative improvement. In average, 12.1% of relative improvement are gained. Using the "Buckeye" AM, only 4.6% of averaged relative improvement were achieved. Higher orders still improve the accuracy but converge to a saturation level.

Analyzing the error rate of the ASR output revealed that mainly utterances with low recognition accuracy can be improved by the proposed approach. To further enhance the performance, we assess the reliability of the ASR output by using the generalized word posterior probability (GWPP) approach [21]. We enumerate various thresholds and send only utterances with lower confidentiality (CF) values than the threshold to the P2W (P2W (CF 50best)). Thereby, in case of "WSJadaptBuck" a relative improvement to the adapted AM of 13.8% could be achieved in average. The 15mix variant performs best and reaches up to 47.8%. In the tests with the "Buckeye" AM, only a relative improvement of 4.9% to the pure spontaneous AM could be achieved. Again, the 15mix variant performs best and achieves up to 62.6% WA. Additionally, the upper bound of our proposed system is shown in Fig. 7 when sending only the utterances to the P2W which can be improved by the system (P2W UpperBound). The upper bounds show that there is only little space left for further improvements (average relative improvement to P2W (CF 50best): 6.7% in case of "WSJadaptBuck" and 5.6% in case of "Buckeye").

The results of Fig. 7 (a) and 7 (b) show that P2W consistently improves the performance on both, the adapted AM and the pure spontaneous AM. The improvement on the adapted AM is higher than the pure spontaneous AM. This may be due to the fact that the pure spontaneous AM already covers a higher amount of pronunciation variation of spontaneous speech than the adapted AM. Therefore, P2W could help more to improve the adapted AM. To further investigate the effectiveness of P2W on the adapted AM, we conducted additional experiments on adapted AM for various word ranges.

Figure 8 shows the performances of "WSJadaptBuck" when increasing the word range. Only the results of the 15mix variants are presented since those AMs performed best in average across the experiments.

Increasing the word range, the WA could still be improved with the proposed approach. However, the improvements start to saturate slightly. The performance saturation for larger word ranges can be reflected in the performance saturation of the proposed approaches when testing on "clean" transcriptions (see Fig. 6 (a)). This may be due to the lack of training data for larger word ranges.

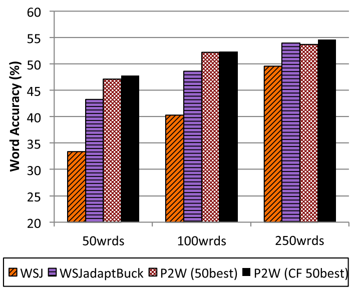
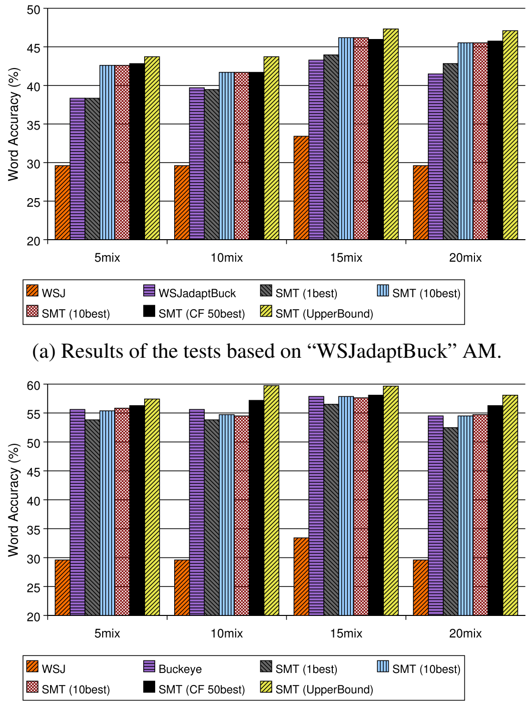


Fig. 8 Results of testing the P2W system on the 15mix variant of "WSJadaptBuck" for all word ranges.



(a) Results of the tests based on "WSJadaptBuck" AM.

(b) Results of the tests based on "Buckeye" AM.

Fig. 9 SMT test results based on the 50words Buckeye test set.

(b) ASR Improvement Using SMT

Similar results have been achieved when applying the second approach. Again, for the 50 words test set the results of all mixture variants (based on the "WSJadaptBuck" and the "Buckeye" AM) are illustrated in Fig. 9.

Given the first best path of the ASR output (SMT (1best)), only the 10mix variant of the adapted AM could not be improved. In case of the pure spontaneous AM, none of the AMs could be improved. Keeping the lattice helps to improve all variants of "WSJadaptBuck" and an averaged relative improvement of 8.2% to the adapted AM is achieved. Here, the 5mix variant achieves with 11.1% the highest relative improvement. As for the "Buckeye" AM, keeping the lattice only helps to improve the 5mix variant.

When the reliability is assessed, using the adapted AM only the 5mix and the 20mix variant could be further improved. The 15mix variant performs best and reaches up

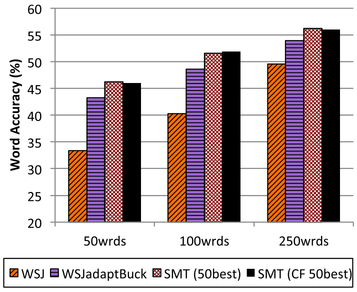


Fig. 10 Results of testing the SMT system on the 15mix variant of "WS-JadaptBuck" for all word ranges.

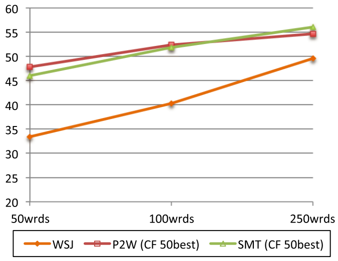


Fig. 11 Comparison of P2W and SMT based on the 15mix variant of 'WSJadapt-Buck".

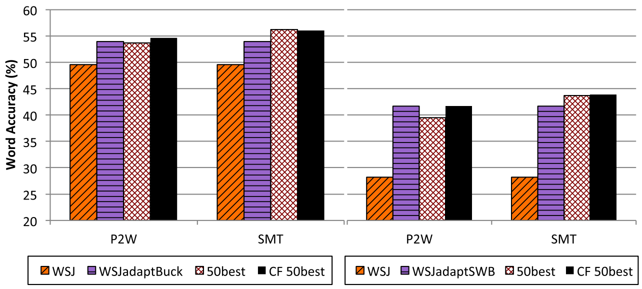


Fig. 12 Comparison of P2W and SMT for both corpora based on the 15mix AMs for the 250words test set.

to 46.0%. By assessing the reliability of the "Buckeye" AM results, finally all mixture variants were improved and an average relative improvement of 2.2% was gained. The 15mix variant performs with 58.1% WA best.

The upper bound shows that there is only little space for improvement (average relative improvement to P2W (CF 50best): 3.2% in case of "WSJadaptBuck" and 3.1% in case of "Buckeye").

Again, only the results of testing the SMT on "WS-JadaptBuck" are presented in detail in the following.

In Fig. 10 the performances of the 15mix "WSJadapt-Buck" AMs are illustrated when the word range is increased. When using SMT, the WA could still be improved for larger word ranges but the improvements start to saturate.

3.5 Discussion

Figure 11 compares the performance results of the P2W and SMT of "WSJadaptBuck". The corresponding baseline result "WSJ" is illustrated as reference.

Both NLP techniques show similar characteristics. In the 50words test set, the P2W achieved higher improvements than the SMT. However, if the word range is increased the improvements start to saturate. However, the performance reduction of the P2W is larger than the SMT. The lower performance for larger test sets is due to the lack of training data.

To prove the reliability of the proposed approach for other data sets, Fig. 12 illustrates the results of P2W and SMT when tested on Buckeye and Switchboard. Here, only the results of the 15mix AMs of the 250words test set are shown. Test results based on Switchboard are generally lower than on Buckeye. When P2W is applied on the 250words test set of Switchboard, the WA could not be improved. In contrast, the SMT achieves to improve the WA for both corpora, even for larger word ranges.

Joint-sequence models cannot deal with deletions and insertions since a mistakingly inserted phoneme cannot be 'ignored" and a missing phoneme cannot be produced by P2W. While training, the order of the N-gram LM can be increased by ramping up the system iteratively. The SMT system is able to deal with insertions and deletions, as it was developed to translate from a source language to a target language where those problems frequently occur. When tests on Switchboard were conducted where the ASR output contained many insertions, it could be revealed that the P2W has difficulties to handle insertions. In contrast, the SMT performed well in those cases.

4. Conclusions

This paper proposes a noisy channel model for modeling pronunciation variation of spontaneous speech. Two approaches (joint-sequence models and statistical machine translation) are derived from the noisy channel model theory and applied in the experiments using the Buckeye and the Switchboard corpus. For both corpora, the performance results show similar characteristics and the results show that both approaches improve the WA consistently over the conventional recognition system. With the Buckeye corpus using P2W an averaged relative improvement of 13.8% to the baseline could be achieved. When SMT was used a relative improvement of 9.3% in average was achieved. Comparing both approaches revealed that the SMT achieves better results for larger word range and performs more robust than the P2W.

The results of this study point towards the positive direction opening the possibility to increase the vocabulary size and the complexity of the experiment's topology. Hybrid approaches which combine the P2W and SMT systems could also be examined.

References

[1] D. Pallet, "A look at NISTS's benchmark ASR tests: Past, present, future," Proc. ASRU, pp.483–488, 2003.

[2] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J.

McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Stochastic pronunciation modelling from handlabelled phonetic corpora," Proc. ETRW on Modeling Pronunciation Variation for Automatic Speech Recognition, pp.109–116, 1998.

[3] T. Jang, "Generation and selection of english pronunciation variations using knowledge-based rules and speech recognition techniques," Studies in Phonetics, Phonology and Morphology, vol.12.2, pp.361–375, 2006.

[4] A. Bates, M. Osterndorf, and R. Wright, "Symbolic phonetic features for modeling of pronunciation variation," Speech Commun., vol.49, pp.83–97, 2007.

[5] K. Chen and M. Hasegawa-Johnson, "Modeling pronunciation variation using artificial neural networks for English spontaneous speech," Proc. ICSLP, pp.1461–1464, 2004.

[6] K. Livescu and J. Glass, "Feature-based pronunciation modeling for speech recognition," Proc. HLT/NAACL, 2004.

[7] S. Sakti, S. Markov, and S. Nakamura, "Probabilistic pronunciation variation model based on Bayesian networks for conversational speech recognition," Second International Symposium on Universal Communication, 2008.

[8] E. Fosler-Lussier, "Contextual word and syllable pronunciation models," Proc. IEEE ASRU Workshop, 1999.

[9] M. Bisani and H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion," Speech Commun., vol.50, pp.434–451, 2008.

[10] P. Koehn, F.J. Och, and D. Marcu, "Statistical phrase-based translation," Proc. ACL/HLT, pp.127–133, 2003.

[11] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," IEICE Trans. Inf. Syst., vol.E87-D, no.8, pp.2121–2129, Aug. 2004.

[12] A. Stolcke, "SRILM - an extensible language modeling toolkit," Proc. ICSLP, pp.901–904, 2002.

[13] F. Och and H. Ney, "A systematic comparison of various statistical alignment models," Computational Linguistics, vol.29, no.1, pp.19–51, 2003.

[14] A. Finch, E. Denoual, H. Okuma, M. Paul, H. Yamamoto, K. Yasuda, R. Zhang, and E. Sumita, "The NICT/ATR speech translation system for IWSLT 2007," Proc. IWSLT, pp.103–110, 2007.

[15] D.B. Paul and J. Baker, "The design for the Wall Street journal-based CSR corpus," Proc. ICSLP, 1992.

[16] S. Pallett, J. Fiscus, M. Fisher, J. Garofolo, B. Lund, and M. Przybocki, "1993 benchmark tests for the ARPA spoken language program," Proc. Spoken Language Technology Workshop, 1994.

[17] P. Mark, K. Johnson, E. Hume, S. Kiesling, and W. Raymond, "The buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability," Speech Commun., vol.45, pp.90–95, 2005.

[18] R. Fasold, The Sociolinguistics of Language, Blackwell Publishers, Oxford, 1990.

[19] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," Proc. ICSLP, pp.24–27, 1996.

[20] S. King, C. Bartels, and J. Bilmers, "Small vocabulary tasks from Switchboard 1," Proc. EUROSPEECH, pp.3385–3388, 2005.

[21] W. Lo and F. Soong, "Generalized posterior probability for minimum error verification of recognized sentences," Proc. ICASSP, pp.85–88, 2005.

Hansjörg Hofmann received the B.Sc. degree in Telecommunications- and Media Technology in 2007 and the M.Sc. degree in Information Systems Technology in 2010 from Ulm University, Germany. Before graduating in 2010, he absolved his Master Thesis at the National Institute of Information and Communication Technology (NICT) in Kyoto, Japan. He is presently pursuing his Ph.D. degree at Ulm University, Ulm, Germany. Now, his research interest are the design of speech interfaces to web services in the automotive environment.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2002, she received her M.Sc. degree in Communications Technology from University of Ulm, Germany. Between 2003-2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006-2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her studies (2005-2008) with the Dialog Systems Group University of Ulm, Germany, and received her Ph.D. degree in 2008. Currently, she is an assistant professor of the Augmented Human Communication Lab, Graduate School of Information Science, Nara Institute of Science and Technology, Japan. Her research interests include statistical pattern recognition, speech translation and graphical modeling framework.

Chiori Hori received the B.E. and the M.E. degrees in electrical and information engineering from Yamagata University, Yonezawa, Japan in 1994 and 1997, respectively. From April 1997 to March 1999, she was a Research Associate at the Faculty of Literature and Social Sciences of Yamagata University. In April 1999, she started the doctoral course in the Graduate School of Information Science and Engineering at Tokyo Institute of Technology (TITECH) and received her Ph.D. degree in March 2002. She was a Researcher in NTT Communication Science Laboratories (CS Labs) at Nippon Telegraph and Telephone Corporation (NTT), Kyoto, Japan from April 2002 to March 2004. She was a Project Researcher at InterACT in Carnegie Mellon University (CMU) in Pittsburgh from April 2004 to March 2006. She is currently a senior researcher at Spoken Language Communication Laboratory at National Institute of Information and Communications Technology (NICT), Kyoto, Japan since 2007. She has received the Awaya Kiyoshi Science Promotion Award from the Acoustical Society of Japan (ASJ) in 2001, Paper Award from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2003, 24th TELECOM System Technology Award from the Telecommunications Advancement Foundation (TAF) in 2009, and International Cooperation Award from the ITU Association of Japan (ITU-AJ) in 2011. She is a member of the IEEE and the ASJ.

Hideki Kashioka received his Ph.D. in Information Science from Osaka University in 1993. From 1993 to 2009, he worked for ATR. From 2006, he works for NICT. He is currently the director of Spoken Language Communication Laboratory at Universal Communication Research Institute, NICT. He is also the visiting associate professor of the graduate school of Information Science at NAIST from 1999.

Satoshi Nakamura received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994-2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000-2008, and Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009-2010. He is ATR Fellow. He is currently Professor of Graduate School of Information Science at Nara Institute of Science and Technology. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics, the Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology, and The Commendation for Informatization Promotion by the Minister of Internal Affairs and Communications.

Wolfgang Minker is a full-time Professor at the University of Ulm, Institute of Communications Engineering (Germany). He received his Ph.D. in Engineering Science from the University of Karlsruhe (Germany) in 1997 and his Ph.D. in Computer Science from the University of Paris-Sud (France) in 1998. He has been researcher at the Laboratoire d'Informatique pour la Mcanique et les Sciences de l'Ingnieur (LIMSI-CNRS), France, from 1993 to 1999 and member of the scientific staff at Daimler-Chrysler, Research and Technology (Germany) from 2000 to 2002.