ヒューマンインタフェース 97−29
音声言語情報処理 40−29
（2002. 2. 2）

Inter-Word Pauses Modeling for Recognizing Noisy Speech in SPINE2 Project

Jin-Song Zhang, Konstantin Markov, Tomoko Matsui, Satoshi Nakamura
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan
{jinsong.zhang, konstantin.markov, tomoko.matsui, satoshi.nakamura}@atr.oc.jp
Tel: (0774) 95-1314

ABSTRACT

The varying background noises make it a problem to model the Inter-Word Pauses (IWPs) in SPINE2 project, easily leading to miss-location of IWPs and ill-estimation their HMMs. This paper presents our approaches to develop explicit acoustic modeling of IWPs and carryout phone-duration-analysis to correctly locate IWPs in the noisy training data. Through iterated optimizations, the final cross-word CD tri-phone HMMs achieved by 9.2% less errors than the initial one trained through a flat-start building procedure. Furthermore, we propose to treat IWP as one word and model it into the language model, this approach successfully reduced the increased computation from acoustic modeling of IWPs, with no significant decrease of the whole recognition performance.

Keywords SPINE project, noisy speech recognition, inter-word pause, duration analysis, prosodic phrase boundary.

SPINE2プロジェクトのための単語間ポーズモデルによる耐雑音性に優れた音声認識

チョウ・キンソン, マルコフ・コンスタンチン, 松井知子, 中村哲
ATR 音声言語コミュニケーション研究所
〒619-0288 京都府相楽郡精華町光台二丁目2番地2
電話：(0774) 95-1314

摘　要

SPINE2プロジェクトで扱う音声データは、逐次的に変化する種々の雑音を含む。それらの雑音により、単語間のポーズ部分を正しく切り出せないために、そのポーズモデルとポーズの左右にある音素モデルを適切に推定することは困難になる。本稿では音響モデルの学習において、単語間ポーズモデルを頑健に作成する手法、および音素の継続時間の分析に基づいてポーズを切り出す手法について報告する。本手法では音響モデルの最適化を複数のステップで行うが、最終的に得られるクロスワード文脈依存トライフォンモデルは、初期モデルと比べて、認識誤り率を9.2％ほど削減する。さらに、単語間ポーズモデルを一つの単語として扱い、言語モデルの中に組み込む方法についても報告する。この方法を用いれば、認識性能をほとんど劣化させることなく、認識に必要な計算量を削減することができる。

キーワード： SPINEプロジェクト、雑音下での音声認識、単語間ポーズ、接続時間分析、韻律句境界

1 Introduction

The second "Speech in Noisy Environments" (SPINE2) evaluation was conducted by the Naval Research Laboratories (NRL) in October 2001. The purpose of the evaluation was to provide continuing forum for assessing the state of the art practice in speech recognition technology for noisy military environments and for exchanging information on innovative speech recognition technology in the context of fully implemented systems that perform realistic tasks. The approach ATR has taken to this task to develop acoustic models is multi-session estimation including robust estimation of baseline HMMs, gender dependent adaptation and channel dependent adaptation, as illustrated in Figure 1.

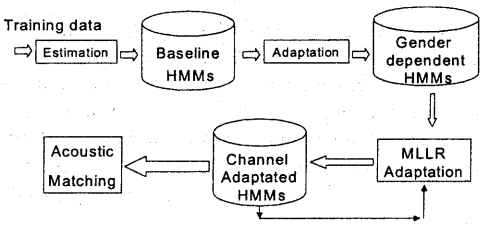

Figure 1: The multi-session adaptive procedure to develop HMMs.

This paper introduces our development of the baseline acoustic models, especially the modeling of inter-word pauses (IWPs), in order to improve the system performance. The main reason for specific considerations about IWPs are that IWPs are very frequent and inhomogeneous in the task, appropriate modeling can not only enhance the robustness of HMMs for themselves, but also improve the accuracy of other phone HMMs. For a description of ATR's whole approach, readers are suggested to look at [1].

The paper is arranged as follows: Section 2 describes the data and shows how frequent the IWPs are in the training data; Section 3 introduces the development of appropriate acoustic modeling of IWPs for speech recognition; Section 4 introduces our another approach to model IWPs by means of language modeling. Finally, section 5 gives conclusions.

2 SPINE2 Data

The SPINE2 data is organized in conversations between two speakers collaborating in a task of seeking and shooting targets. Each speaker is seated in a different sound chamber in which previously recorded background noise environment is reproduced. Push-to-talk recordings were made of signals from a communication line. The line was activated at approximately the time at which an utterance began, and was deactivated at the end of the utterance. The recorded signal consisted of a continuous background signal of noise produced by the recording equipment, with intermittent recordings of the speech and reproduced noise communicated through the channel [2]. There are total of 11 types of noisy environments including quiet, office, aircraft carrier, street, car, helicopter, tank, fighter jet and others. Besides the noise background, there are also sounds of whistles, rings, additional tones, background speech etc. Additionally, the speakers talked freely so that dropouts, repairs and other kinds of spontaneous speech phenomena are also frequent.

2.1 Training and Testing Data

Training data consists of 324 dialogs involving 20 speakers (10 males and 10 females). There are about 28000 utterances with average length of 4 seconds. Total duration of speech data for training is about 15 hours. The signal-to-noise ratio (SNR) varies from 5dB to 20dB. All data have only transcripts at word level, no phonetic segmentation information.

As test data, we used 8 channels of 4 conversations from the development data, between 2 male and 2 female talkers who are different from the training speakers, with the following four noise environments: quiet, office, helo(helicopter) and bradley (tank), 2 channels each. The total number of utterances is 361.

For the signal processing, although we have used different feature representations in our system [1], all the experiments here used the standard mel-scale cepstrum (MFCC). Speech data were sampled at 16kHz and frame size of 20ms and frame shift of 10 ms were used to compute MFCCs. 12 MFCCs plus log energy and their 1st and 2nd order time derivatives form a 39 dimensional vector.

2.2 Language Model

The language model (LM) training data and task vocabulary were provided by CMU and were common for all participants in this SPINE2 evaluation. Using the training data, we developed word bigram language model for all the experiments here. The training data also contains the transcripts of the development data. During the experiments, the language model scale was fixed to the same value (7) in order to clarify the effects from different acoustic models.

2.3 Inter-word Pauses (IWPs)

It was noted that silent pause segments are very frequent in the SPINE2 data. Figure 2 illustrates the frequency histogram for the pauses and the top-10 most frequent phones in the training data, collected from the phonetic segmentation aligned based on our final acoustic models AM09 (described in next section).

- *ps0* stands for a silence in the beginning of an utterance.

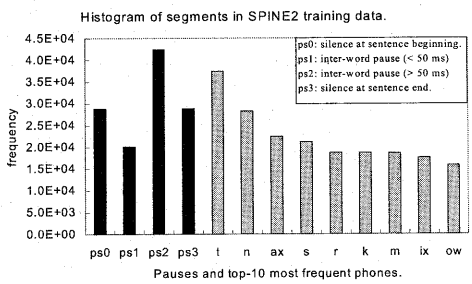

Figure 2: Histogram for the pauses and top-10 most frequent phones in the SPINE2 training data.

- *ps1* stands for an inter-word silent pause whose duration is longer than 10ms but shorter than 50ms.
- *ps2* stands for an inter-word silent pause whose duration is longer than 50ms.
- *ps3* stands for a silence in the end of an utterance.

It shows that not only pauses are very frequent but also the long inter-word pauses (*ps2*) are even the most frequent one in the data. The reason for this is that most utterances in the task consist of series of military commands. The inherent prosody structures own frequent concatenations of short command phrases. Hence, pauses, which are usually associated with intonation phrase boundaries, are very frequent in this task. Due to the push-to-talk data collection method and various background noises, the IWPs probably have different statistics from those silent segments in two utterance-ends. Therefore, appropriate modeling for the pauses should be necessary in order to improve the recognition performance.

3 Acoustic Modeling of IWPs

Acoustic modeling of IWPs includes two aspects in a speech recognizer based on context dependent HMMs: appropriate HMM for IWPs themselves and appropriate modeling of their contextual effects on their neighboring phones. The most known method is the *sp* (short pause) tee HMM given in HTK book [3], and the key points are:

- The *sp* tee HMM has only one skippable state which is tied to the center state of a 3-state *sil* HMM which is for the silences at both utterance-ends.
- Each word pronunciation is attached by *sp* to model any possibly-appearing IWPs in speech.
- The *sp* is *Context Free*: it does not block context-dependent effects in a cross-word context modeling system.

This approach typically brings about good performances for recognition of clean speech, and it serves as our starting point. However, there are two possible questions associated with this approach when applied to SPINE2 task.

1. IWPs in SPINE2 probably are different from those silences at utterance-ends. The direct tying of *sp* and *sil* may not be appropriate.
2. IWPs in SPINE2 may be long enough to block the coarticulation effects between two neighboring phones, suggesting that they should not be *Context Free*.

In order to make clear these doubts and testify any new proposals, we carried out a series of studies on the effects of different modeling of IWPs. Finally, we reduced the word error rates by absolute 9.2% from 53.21% to 44.0% of the cross-word CD tri-phone HMMs using MFCC features and the same LM scale.

Flat-start-HMM building: Since phonetic segmentation was not available initially for the training data, we need to develop HMMs from flat-start, and do estimations, re-alignments and optimizations by a number of iterations [3]. All the experiments in this paper used this method to develop HMMs.

3.1 *sp* HMM's Estimation

The first investigation was made to the effects of untying of *sp* and *sil* HMMs in 3 different ways.

- Method 1: *sp* was initialized from the "*sil*" HMM after flat-start mono-phone estimations, and also tied to *sil* [3]. This is the conventional way.
- Method 2: *sp* was initialized in the same way as Method 1, but not tied to *sil*. If the assumption that IWPs are statistically different from the silences at utterance-ends, this method should be better than Method 1.
- Method 3: *sp* was initialized in the same way as other flat-start mono-phones, and not tied to *sil*. This method offers the most freedom for estimation of *sp*, assuming no relation between *sp* and *sil*.

	Method 1	Method 2	Method 3
WER %	53.8	52.6	51.6

Table 1: Results in word-error-rates (WER) of different estimations of *sp*.

The HMMs for other phones in this experiment was state-tied intra-word tri-phones, each with left-to-right 3 state topology. 8460 tri-phones from the lexicon share 2000 state, each with 13 Gaussian mixtures. Table 1 gives the recognition results for

the test data with respect to the 3 methods to estimate *sp*. The results showed that Method 1 got the most errors, and Method 2 achieved 1.2% less, and Method 3 achieved the least. Therefore, the results provide evidence for our assumption that IWPs might have different distributions from silences at utterance-ends, and separate modeling of IWPs and *sil* may lead to better performance.

3.2 Different Context Modeling

The second investigation was made to effects of different context dependent (CD) modeling. We developed the following 4 sets of acoustic models:

- AM01: intra-word CD tri-phone HMMs.
- AM02: cross-word left CD di-phone HMMs.
- AM03: cross-word right CD di-phone HMMs.
- AM04: cross-word CD tri-phone HMMs.

The four models estimated *sp* in the same way as the previous Method 2, having the similar number of tied states and Gaussian mixtures. Figure 3 shows the recognition results and the average training samples per allophone (ASP). Observations about the results suggest:

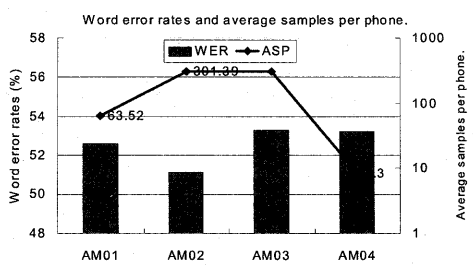

Figure 3: The relations between word error rates (WER) and average training samples per allophone (ASP) of different context modeling.

1. Although cross-word CD tri-phone modeling is assumed to be more powerful than intra-word CD tri-phones, AM04 got more errors than AM01. But we cannot conclude cross-word CD modeling is not effective in this task. Because AM02, the cross-word CD di-phone achieved less errors than AM01.

2. The probable reason may be attributed to the factor of different ASPs of each modeling. Although each tied state in the different models may share similar number of training data, the state-tying procedure is dependent on the acoustic estimations of the allophone HMMs. When the ASP is few, as only 8.3 per phone in the case of AM04, it is difficult to get robust allophone HMMs. Hence not easy to achieve robust state tying of HMMs.

3. Furthermore, the coarticulation-block effects by the frequent IWPs were not considered either in AM04. For example, a phone sequence "*A long-pause B*" may be accurately modeled in the way of "*A+{pause} pause {pause}-B*" from the view of coarticulation. However, in the conventional way of *Context Free* modeling of IWPs such as *sp*, they are modeled by "*A+{B} sp {A}-B*". Ignorance of the correct modeling of IWPs lead to possible ill-estimation of the allophone HMMs.

4. The lower performance of AM03 than AM02 can be attributed to the modeling of different coarticulation effects. Left CD di-phone models the carryover effects, which are phonetically more significant than the anticipation effects modeled by the right CD di-phones. Comparison of the two models also showed that ASP is not a dominant factor for developing robust HMMs.

3.3 Explicitly Modeling IWPs

Based on the analyses of the two preliminary investigations, it is obviously reasonable to adopt a new symbol *ps* to explicitly model IWPs, and its HMM was developed in the following way:

- The *ps* HMM has the same topology as the *sil*, and was initialized by the *sil* HMM after flat-start mono-phone initialization.
- The *ps* is *Context Dependent*, appearing in the context factor of CD modeling.
- A separate dictionary entry with *ps* suffix is created for each real pronunciation in the lexicon.
- Initial training samples for *ps* took those *sp* segments whose durations were longer than 50ms. Then iterated force-alignments of the training data automatically found the phonetic segmentations in the later training produres.

Based on the procedure, we developed a new set of acoustic model AM05:

- AM05: Cross-word CD di-phone HMMs with a *ps* to explicitly model IWPs.

The recognition performance of AM05 is given in Figure 4. When compared to AM02, AM05 reduced WER by 2.6% from 51.1% to 48.5%, showing the effectiveness of the proposal.

3.4 Duration Analysis Based location of IWPs

When we paid a look at the phonetic segmentations resulted from force-alignment by AM05, we found those IWPs with high noise level were frequently miss-located by either *ps* or *sp*. Their statistics could not be learned by either *ps* or *sp*. In order to locate these noisy IWPs, we developed a duration analysis based approach.

- Step 1: compute the duration mean μ_i and deviation σ_i of each phone P_i from the phonetic segmentations of the training data aligned by AM05.
- Step 2: If a word boundary phone P_i was not followed or preceded by an IWP label, and its duration is extraordinarily long ($> \mu_i + 3 \times \sigma_i$), an *ps* label would be inserted for a miss-located IWP.

The philosophy under this approach is that a phrase-final vowel is often lengthened and the duration of a consonant is relatively consistent [5]. It is reasonable to assume a *pause* following an extraordinarily long phrase boundary vowel, and regard an extraordinarily long consonant as including a miss-located *pause*. With the new phonetic segmentations, we trained another set of HMMs.

- AM06: Cross-word CD di-phone HMMs with *ps*, estimated from duration-based located I-WPs.

Figure 4 showed that AM06 further reduced WER by 1.7% from 48.5% of AM05 to 46.8%, demonstrating the positive effect of the proposal.

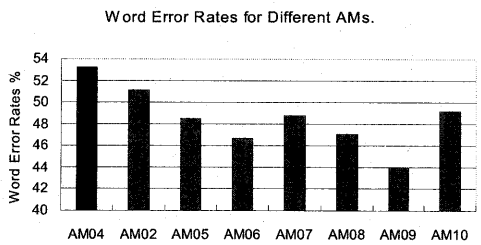

Figure 4: Recognition results of different acoustic models.

3.5 A HMM for Noisy IWPs

We also tried a separate *np* HMM to intentionally model noisy IWPs. The *np* HMM was developed in the same way as *ps*. The initialization samples for *np* took those *ps* segments in noisy channels, denoted by provided channel information. The acoustic model with *np* is AM07.

- AM07: Cross-word di-phone HMMs with *ps* and *np*.

However, this approach led to by 2.1% more errors than AM06. And the reason was analyzed as: both *ps* and *np* are *Context Dependent*. The resulting multiple CD di-phones of *ps* and *np* may be biased to the channel background noises, unable to robustly estimate an HMM for IWPs.

3.6 Merge Silent HMMs

By now, we have already 4 types of silent HMMs: a *sil* for silences at utterance-ends, three HMMs, i.e., *sp, ps, np*, for IWPs. As shown in the previous experiment, an increasing number of silence HMMs had the possible problem of insufficient estimation of the parameters. From the view of coarticulation, long IWPs have the similar contextual effects on their neighboring phones to those arouse by silences at utterances-ends. Therefore, it should be reasonable to merge the *ps* and *np* to *sil* in order to get a robust estimate of all silences. Since we have got more accurate phonetic segmentations of the training data from the previously evolving HMMs like AM06 or AM07, we may rely on the segmentations of IWPs to estimate *sil* rather than on the on-line alignment of training data. Another 3 sets of acoustic models were developed.

- AM08: Cross-word CD di-phone HMMs with the *sil* for IWPs either. Phonetic segmentation was made based on AM07.
- AM09: Cross-word CD tri-phone HMMs with the *sil* for IWPs either. AM07 based phonetic segmentation was used.
- AM10: Cross-word CD tri-phone HMMs with the *sil* for IWPs either. Phonetic segmentations was from on-line alignment, rather than the one segmented by AM07. This is similar to the method introduced in [4].

From Figure 4, we may see that:

1. AM08 achieved by 1.7% less errors than AM07, but still 0.4% more errors than AM06. It may be ascribed to the difference between *Context Independent* modeling of IWPs by *sil* and *Context Dependent* modeling by *ps*. The previous modeling owns substantially less number of parameters.
2. The shortcoming of AM08 was overcomed in AM09 by adopting the tri-phone modeling which substantially increased the number of allophones. The accurate phonetic segmentation by AM07 guaranteed a robust estimation and tying of the allophone HMMs. This resulted in the lowest error rates, by a further 2.7% when compared to AM06, and by 9.2% when compared to the initial cross-word CD tri-phone AM04.
3. As a comparison experiment to AM09, AM10 got 5.2% more errors. This indicates the significant effect of correct phonetic segmentations of IWPs on the development of cross-word CD tri-phone HMMs.

4 Language Modeling of IWPs

The previous acoustic modeling of IWPs need to adopt optional pause suffices to each real pronunciation in the lexicon. This means that the adoption of one model *ps* would double the size of the original lexicon, and adoption of two *ps and np* would

make the new lexicon 3 times large as the original one. As a result, decoding computation also increased by several times.

One substitute way is to regard the IWP as one word, and develop language model from speech aligned transcripts with located IWPs. Then the probabilities such as:

$$\text{Prob}(IWP|W_{i-1}W_{i-2}\ldots,W_{i-n+1})$$

$$\text{Prob}(W_i|W_{i-1}\ldots,IWP\ldots,W_{i-n+1})$$

can be estimated to model any possible statistical relation between the normal words and IWP word. Based on the new language model, not only the computation increased by modeling IWPs can be reduced, but also the new language model has the ability to model the underlying prosody structure. As IWPs are likely to follow only those words which can appear as phrase finals[6].

After aligning all the training data and development data based on AM07, we got new word transcripts with IWPs, and used them to estimate a new word bigram language model. The new language model has a better perplexity than the original one, as in Table 2.

Language Model	Perplexity
LM1 (normal)	16.3
LM2 (IWP)	15.6

Table 2: Perplexity of two language models.

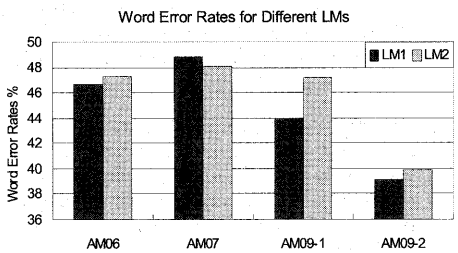

Figure 5: The effects from different LMs on recognition results.

Based on the new language model LM2, we carried out recognition experiments using acoustic models AM06, AM07 and AM09. The results are given in Figure 5, where AM09-1 used the same LM scale 7 as before, and AM09-2 used an optimized LM scale 15, which is the best one for each LM in most cases.

The results showed that LM2 introduced a little more errors than LM1 in most cases. Although LM2 got by 3.2% more errors than LM1 for AM09 when using the fixed LM scale 7, the gap was reduced to 0.8% after LM scale was optimized. From these results, we suggest:

1. It is reasonable to model IWPs by language models, as evidenced by the better perplexity of LM2, and comparable recognition performances of LM1 and LM2 with respect to different acoustic models.
2. High-order n-gram LM should be experimented to draw a more general conclusion.
3. The advantage of reduced computation by LM2 should be notified.

5 Conclusions

This paper discussed the problem of IWPs' modeling in SPINE2 task, which arouse from the different acoustics of IWPs from the silences at utterance-ends, and introduced our approach to exploit either acoustic modeling or language modeling to deal with the problem. Based on iterated optimization of the phonetic segmentation of IWPs, the final cross-word CD tri-phone HMMs achieved by 9.2% less errors than the initial one trained through a flat-start building procedure. The language model approach successfully reduced the increased computation from acoustic modeling of IWPs, with no significant decrease of the whole recognition performance.

Acknowledgements We would like to thank Mr. Rainer Gruhn, Mr. Norbert Binder and Mr. Masaki Ida for their help and valuable discussions during the task.

References

[1] K. Markov, T. Matsui, R. Gruhn, J.-S. Zhang, and S. Nakamura, "ATR system for Robust speech recognition in real world noisy and channel environments - Evaluation on DARPA SPINE2 task", Proc. of IEICE and ASJ symposium on SLP, Dec., 2001, Tokyo.

[2] R. Singh, M.L. Seltzer, B. Raj and R. M. Stern, "Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination", Proc. of ICASSP2001, Salt Lake City.

[3] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev and Ph. Woodland, "HTK Book: version 2.2"

[4] T. Hain, P.C. Woodland, G. Evermann and D. Povey, "The CU-HTK March 2000 HUB5E Transcription system", Proc. of DARPA 2000 Speech Transcription Workshop.

[5] R. D. Kent and Ch. Read, "The acoustic analysis of speech", Singular publishing Group, Inc., 1992.

[6] H. Niemann, E. Noth, A. Batliner, J. Buckow, F. Gallwitz, R. Huber, A. Kiebling, R. Kompe, V. Warnke, "Using prosodic cues in spoken dialog systems", Proc. of SPECOM'98, ST-Petersburg, 1998, pp.17-28.