PAPER *Special Issue on Speech Information Processing*

A Study on Acoustic Modeling of Pauses for Recognizing Noisy Conversational Speech

Jin-Song ZHANG[†], **Konstantin MARKOV**[†], ***Nonmembers***, **Tomoko MATSUI**[†], ***and*** **Satoshi NAKAMURA**[†], ***Members***

SUMMARY This paper presents a study on modeling inter-word pauses to improve the robustness of acoustic models for recognizing noisy conversational speech. When precise contextual modeling is used for pauses, the frequent appearances and varying acoustics of pauses in noisy conversational speech make it a problem to automatically generate an accurate phonetic transcription of the training data for developing robust acoustic models. This paper presents a proposal to exploit the reliable phonetic heuristics of pauses in speech to aid the detection of varying pauses. Based on it, a stepwise approach to optimize pause HMMs was applied to the data of the DARPA SPINE2 project, and more correct phonetic transcription was achieved. The cross-word triphone HMMs developed using this method got an absolute 9.2% word error reduction when compared to the conventional method with only *context free* modeling of pauses. For the same pause modeling method, the use of the optimized phonetic segmentation brought about an absolute 5.2% improvements.

key words: *Conversational speech recognition, noisy speech, Hidden Markov Model, inter-word pause, context dependent HMMs, duration analysis, prosodic phrase boundary.*

1. Introduction

Normal speech flow usually includes a number of silent periods [1], including silences at utterance-ends, inter-word pauses, and intra-segmental pauses like the voice onset time of a stop consonant. The existence of silences and intra-segmental pauses is relatively more stable than that of inter-word pauses, as they can be reliably inferred from word transcripts of speech data, while the inter-word pauses (hereinafter pauses only) have very flexible appearances. Previous studies showed that appropriate modeling of the pauses might improve recognition performance. In [2], the authors proposed to use the length information of pauses to develop phonetic decision tree based tied-state cross-word triphone HMMs and achieved about a 5% relative error reduction compared with those ignoring the pauses. In [3], three types of different word-end pauses were adopted as pronunciation variations for each dictionary entry and the approach successfully led to more than 1% absolute error reduction in a number of tests.

These studies have one common point in that the speech data is almost clean, where pauses have rather different stationary acoustics from the normal speech segments and can be automatically segmented out via iterated forced alignments with the evolved acoustic models. However, speech data from realistic applications may have varying background noises, thus the pauses are contaminated with varying acoustics. This will make it difficult to generate the correct phonetic transcriptions for the pauses, because an initial simple pause HMM won't be able to segment out those varying pauses, and these miss-segmentations will result in a poor estimation of the pause HMM in a later training stage. Then the poorly estimated pause HMM will furthe miss-segment those varying pauses in a later iteration of forced alignment. This kind of circle will finally result in incorrect phonetic transcriptions and poorly estimated acoustic models.

Furthermore, if the speech data is of conversational speaking style, there will be very frequent pauses due to a heavy load of planning speech for speakers and cognition for hearers in conversations [1], [5]. The problem of miss-segmentations of the pauses may lead to significant influences on the acoustic models. Therefore, studies must be made on the problem of how to model and segment varying pauses when developing acoustic models to recognize conversational speech in varying noisy environments.

Instead of integrating the phonetic segmentation process into the development of acoustic models, as is done in conventional approaches, we took the segmentation as a separate stage. This enables us to adopt different modeling for the varying pauses in the segmentation stage from the one used for training the final acoustic models. We propose that the phonetic heuristics of pauses, including coarticulation and prosody effects, can be exploited to robustly initialize the pause HMMs. Such initialized HMMs lead to better phonetic segmentations and reestimated HMMs. After the optimized transcription becomes available, it can be used to train a set of more robust acoustic models.

Studies have been made on the data of the second "Speech in Noisy Environments Evaluation " (SPINE2) task [6], which are conversations in real military varying noises. The segmentation approach, which exploits the pauses' heuristics, is realized as a stepwise optimization approach for the pause HMMs and the phonetic transcriptions. Experimental results showed that the

Manuscript received June 30, 2002.
Manuscript revised November 12, 2002.

[†]The authors are with ATR Spoken Language Translation Communication Laboratories Inc., Kyoto-fu, 619-0288 Japan.

approach effectively detected an increasing number of noisy pauses, and that the final cross-word triphone HMMs based on the optimized phonetic transcription achieved significantly less word errors than the baseline HMMs.

The paper is arranged as follows: Section 2 describes the precise contextual modeling of pauses and its implementation problem when applied to noisy speech recognition. Section 3 introduces the proposal to exploit phonetic heuristics about pauses for collecting initialization samples for pause HMMs. Section 4 introduces the SPINE2 data and experimental set-up. Section 5 presents the experimental results and discussions. Finally, section 6 gives a conclusion.

2. Precise Contextual Modeling of Pauses And Its Implementation Problem

Context dependent (CD) hidden Markov models (HMMs) have been widely used in current large-vocabulary continuous speech recognition (LVCSR) systems. Although the contextual modeling of normal phones is generally the same for representative LVCSR systems such as HTK [4] and JULIUS [7], that of pauses may be specific to a particular system. For example, the standard HTK system treats cross-pause coarticulation as always possible when cross-word triphone dependency is used. On the contrary, the JULIUS system uses a *context independent* pause model, which always shows contextual blocking effects. However, we regard this kind of particular contextual modeling of pauses as insufficient, as phonetic studies have revealed that complex contextual effects might be associated with pauses on the neighboring word boundary phones [8], [10], [12].

2.1 Complex Contextual Effects of Pauses

The first kind of contextual effect of pauses is to block the coarticulation of the two word boundary phones adjacent to it. When a pause is long or significant enough, like one appearing as a boundary between two prosodic phrases, the phone preceding (succeeding) the pause may have special independent articulation patterns[8], [10], showing no or little coarticulation with the phone succeeding (preceding) the pause.

The second kind of contextual effect of pauses is no-influence on the coarticulation of the two neighbouring phones. When a pause is relatively short or when the articulators involved for the two phones have slow damping characteristics[13], the pause may have no or little influence on the coarticulation.

The third kind of contextual effect is that the cross-pause coordination of the articulators is variable [10]. Given the same length of pauses, gestures with articulators of slow damping speed show stronger coarticulation transition trajectories than those with fast articulators. Therefore, the cross-pause coarticulation may be gradient rather than categorically yes or no.

2.2 Accurate Cross-word Triphone Model Training Based on Contextual Free and Influential Pauses

The above phonetic review suggests that the contextual effects of pauses should be cautiously considered in order to train more accurate and more robust cross-word triphone acoustic models. One practical way to do this is to use several pause HMMs with different contextual effects to precisely model the specific coarticulation of a pause [11]. Here, the following two kinds of pauses are adopted.

- *Context Free:* the pauses do not or slightly affect the coarticulation across it. The context free pauses are modeled by the *"sp"* HMM as in [4]. In the means of triphone modeling, the *sp* won't affect the context expansion of its neighboring triphones. As an example, given a training speech of phone sequence *"A B pause C D"*, the triphone expansion for *"B pause C"* would become as: *"A-B+C sp B-C+D"*, if the *pause* is regarded as context free.
- *Context Influential:* the pauses approximately or completely block the coarticulation across them. Such a pause appears in the context of its neighboring triphones to indicate its context influential effects. When using a *"sil"* HMM for these pauses, the triphones for the *"B pause C"* in the above example would become as *"A-B+sil sil sil-C+D"*.

Therefore, an accurate phonetic labeling of the training data, including correct location of pauses and specification of their contextual effects, is important for trainning robust acoustic models. Since it is too expensive and time consuming to manually label the training data, a conventional way for identifying pronunciation variations, referred to as *force alignment* in [4], can be used to identify the pauses and their contextual effects in the training data for training the final acoustic models. First, a set of HMMs can be trained using a phonetic labeling with abitrary pauses. Then an HMM network like in Fig. 1 can be constructed for the example utterance *"A B pause C D"*. Acoustic decoding finds the path with the maximum likelihood, which specifies simultaneously both the existence of pauses

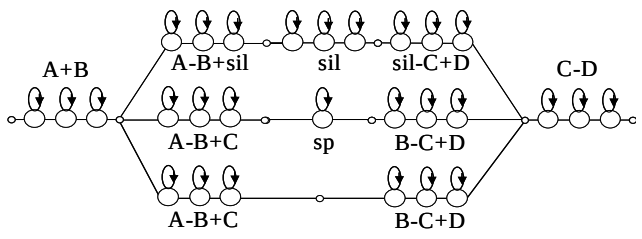

Fig. 1 An example HMM network for decoding during force alignment.

and their contextual effects. Then the renewed phonetic segmentation can be used to estimate the HMMs again. The above processes of model training and force alignment can be realized in a number of iterations in order to get increasingly improved HMMs and phonetic segmentations.

2.3 Pause Identification Problem for Noisy Conversational Speech

The pause HMMs are usually initialized using the silences at utterance starts and ends at the beginning of force alignment. When the training data has a stationary background environment, inter-word pauses have the same acoustics as the silences. The pause HMMs initialized in this way can reliably detect out the optional pauses.

However, when the training data is realistic conversation speech with varying background noises, the pause HMMs initialized from silence data may make it difficult to identify those pauses that have different acoustics from silences. When the pauses are not detected out, their acoustics cannot be learned by the HMMs trained in a later iteration. Then the renewed HMMs still cannot identify those pauses as they donot have their statistics. Any further iterations of model estimation and force alignment cannot solve this problem of inherent miss-identification, and will finally affect the HMMs developed.

The reason can be explained with the aid of Fig. 1. If the pause and its contextual effects are miss-identified, then the utterance will be wrongly assigned to estimate the HMMs of not only a different pause but also two other different neighboring triphones. Since conversational speech inherently owns a large number of pauses, and if such miss-identifications are rather frequent due to severe varying noises, then the portion of trainning data wrongly assigned may be significant enough to result in a set of badly estimated HMMs.

3. Phonetic-Heuristics-Originated Pause Modeling

The pause identification problem can be ascribed to the poor initialization of pause HMMs at the beginning of force alignment. There they are estimated using only silences at utterance starts and ends. If we can get enough samples of noisy pauses for initializing the pause HMMs, the iterative model estimation and force alignment may lead to better HMMs and phonetic segmentations. For this purpose, we propose to exploit two reliable phonetic heuristics of pauses to find sufficient initialization samples for pause HMMs.

3.1 Reliable Phonetic Characteristics of Pauses

The first important characteristic of pauses exploited is the relations between the length of a pause and the phone coarticulation across it [9], [12]. Although the relations are rather complex, depending on not only the length of the pause but also the articulatory configurations of the phones and other factors, [12] showed that a pause longer than 60ms promotes the preservation of distinctive features of consonants at the boundary. This suggests that 50-70ms be a reasonable length limit to assume the context effect of a pause as either *context free* or *context influential* in the initialization stage.

The second characteristic of pauses helpful for their detection is their important relations to the prosodic phrasing boundaries. This is because higher-level phrase boundaries are also known to give rise to longer articulatory durations to the last phones before them, i.e., the phrase-final lengthening, and less influence on the durations of the following phones. This suggests that a pause would probably appear after or before an extra-ordinarily long phone at a word boundary.

3.2 Reliable Initialization of Pause HMMs

Based on these two characteristics, we propose the following methods to find initialization samples for pause HMMs.

1. *Initial phonetic transcription generation:* use the conventional iterated forced alignments to get a phonetic transcription S for the training data.
2. *Pause length based contextual effect specification:* the pauses longer than 50ms in the transcription S are deliberately assigned as *context influential*, and those shorter than 50ms as *context free*.
3. *Phrase boundary pause insertion:* extra *context influential* pauses can be inserted to the transcription S at places where prosody phrase boundaries are assumed to exist, based on statistical phone duration analysis. The method is:
   - First, compute the duration mean μ_i and deviation σ_i of each monophone P_i in S.
   - Second, if a word boundary phone P_i was not followed or preceded by a *context influential* pause, and its duration is extraordinarily long ($> \mu_i + (2 \sim 3) \times \sigma_i$), a *context influential* pause label would be inserted for a possibly miss-located pause.

4. SPINE2 Data and Experimental Set-up

The above proposals have been applied to train acoustic models for the SPINE2 evaluations [11], [14], which consists of spontaneous conversations between pairs of talkers working on a collaborative, battleship-like task. Each person is seated in a sound chamber in which a previously recorded military background noise environment is accurately reproduced. The participants use

the microphone and headset that are resident to the particular environment. There are 11 types of military noisy environments, including quiet, office, HMMWV, aircraft carrier, AWACS, MCE, Bradley tank, car, F16, helo (helicopter) and street. The noises may also be played at varying amplitudes. Part of the data was recorded in a push-to-talk method, resulting in approximately simultaneous appearances of speech and noise signals [15]. Contray to the complexities of background environments, the total vocabulary used is fairly limited.

Pauses are noted to appear very frequently in the data, possibly due to the fact that most of the conversations are series of short military commands, each associated with an intonation phrase boundary. They are contaminated by noises with varying types and varying amplitudes. Furthermore, in the data of push-to-talk recorded, pauses are varying noises while the silences at utterance-ends are clean.

4.1 Training and Testing Data

Training data [14] consists of 628 channels (dialog sides) of 324 dialogs involving 20 speakers (10 males and 10 females). There are about 28,000 utterances with average length of 4 seconds. The total duration of speech data for training is about 15 hours. The signal-to-noise ratio (SNR) varies from 5 dB to 20 dB in the noisy channels [15]. All data have only transcripts at word level, with 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. The channel based average SNRs range from 7.6dB of Helo to 23.9dB of quiet ones.

Table 1 and Table 2 summarize the noisy environments in the training data and testing data respectively. One thing to be noted is that although the four kinds of noisy environments, i.e., quiet, office, Bradley and the helo, shared balanced proportions of the testing data, they have different amounts of training data. The Bradley and the helo environments each have only 16 channels of the training data, far less than those of the quiet and the office ones. So the task is a seriously mismatched testing.

4.2 Experimental Set-up

The acoustic feature used in this study is the standard mel-scale cepstrum (MFCC), as specified in Table 3. The basic lexicon consists of about 5.7k unique words with a total of about 11k entries for pronunciation variations. The basic phone set has 43 American phones, an *sp* for short pauses and a *sil* for pauses and silences at utterance-ends. All the phone HMMs have 3 left-to-right states, except that the *sp* has only one skippable state. The language model used here is a word bi-gram model trained from the transcripts of both training and development data. During the recognition experiments, the language model scale was fixed to the same value in order to clarify the effects from different acoustic models.

Table 1 A summary of the noisy environments in the training data. Each dialogue lasts from 3 to 5 minutes, and is recorded into two channels. The "Percentage" column indicates the proportion of the respective noisy data to the whole training data.

Noise type	# of channels	Percentage (%)
Quiet	176	27.2
Office	136	21.0
HMMWV (vehicle)	100	15.4
Aircraft carrier	76	11.7
AWACS plane	40	6.2
MCE field shelter	40	6.2
Bradley tank	16	2.5
F16 jet fighter	16	2.5
Car	16	2.5
Helo (helicopter)	16	2.5
Street	16	2.5

Table 2 The noisy environments in the testing data used in this study.

Noise type	# of channels	SNR	Percentage (%)
Quiet	2	23.9	25
Office	2	20.8	25
Bradley tank	2	15.8	25
Helo (helicopter)	2	7.6	25

Table 3 Specification of the acoustic feature extraction.

Parameter	Value
Sampling rate	16000 Hz
Frame shift	10ms
Frame length	20ms
Pre-emphasis coef.	0.97
Parameters	12 MFCC + 1 log energy + 13 1st order and 13 2nd order derivatives

5. Experiments

Due to the rather complex variations in the SPINE data, a stepwise procedure based on the previous proposals was used to optimize the initialization and trainning of the pause HMMs, and achieve better phonetic segmentation, after a preliminary investigation to choose a robust kind of context dependent (CD) HMMs for segmentation.

5.1 Choose Robust HMMs for Segmentation

Four sets of different CD HMMs were developed, each having about 2,000 phonetic decision tree based tied states and 16 mixtures of Gaussians per state.

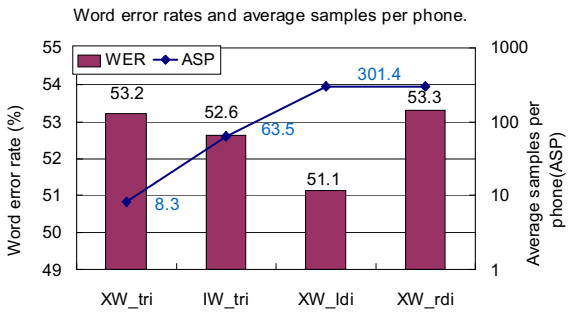

Fig. 2 Word error rates (WER) and average training samples per allophone (ASP) of different context modeling.

- *XW_tri*: cross-word CD tri-phones.
- *IW_tri:* intra-word CD tri-phones.
- *XW_ldi:* cross-word left CD di-phones.
- *XW_rdi:* cross-word right CD di-phones.

The four models used the *sp* for pauses and the *sil* for silences at utterance-ends, as in [4]. Fig. 2 illustrates the recognition performance in word error rates (WER). Observations about the results suggest:

1. Although cross-word CD tri-phone modeling is assumed to be the most powerful among the four kinds of HMMs, it nearly got the highest word error rate, with even more errors than the intra-word CD tri-phones. This is different from usually reported results [4], [16].
2. The best performance was achieved by the model *XW_ldi*, the cross-word left CD di-phone HMMs, with the highest robustness here.
3. The reason for these results might be attributed to the poor modeling of pauses. As only contextual free *sp* HMM is used, inappropriate triphone labels may be assigned to those phones adjacent to pauses. In such a case, a context modeling with higher ASP (average training samples per allophone) should have higher robustness. Therefore, *XW_ldi* has better robustness than those with lower ASP values.
4. The better performance of *XW_ldi* than *XW_rdi* may reflect the fact that carryover coarticulations are more significant than the anticipations in natural speech.

Therefore, cross-word left CD diphone modeling was chosen to be used in the later steps to optimize phonetic segmentations. The model *XW_ldi* generated a phonetic transcription S_{XW_ldi}.

5.2 Stepwise Optimization of Pause Modeling and Phonetic Segmentation

Next, the following optimization steps were realized.

- Step 1: Incorporate a *ps* HMM for pauses to model the phenomenon that pauses may be varying from silences. The *ps* HMM is *context influential*, and its initialization samples took those *sp* segments in S_{XW_ldi} whose durations were longer than 50ms. Through the iterations of model estimations and forced alignments, we developed a new set of cross-word left CD diphone HMMs *XW_ldi_ps* and got a new phonetic transcription $S_{XW\_ldi\_ps}$.
- Step 2: Insert *ps* labels into $S_{XW\_ldi\_ps}$ according to the method of *phrase boundary pause insertion* in order to generate initialization samples for those pauses with high-level noises. Then, through the same iterations as Step 1, we got new HMMs *XW_ldi_dur* and transcription $S_{XW\_ldi\_dur}$.
- Step 3: Incorporate another *np* HMM as a more detailed noisy pause model. The *np* HMM is also *context influential*, and its initialization samples took those *ps* in utterances from noisy channels based on $S_{XW\_ldi\_dur}$. Similarly, we got the model *XW_ldi_np* and the phonetic transcription $S_{XW\_ldi\_np}$.

In each step, speech recognition experiments were carried out based on the developed HMMs in order to show their efficiency.

5.3 Developing Final Acoustic Models

The phonetic transcription $S_{XW\_ldi\_np}$ is assumed to contain a rather correct segmentation of pauses. Then we replace all the labels of *ps* and *np* in $S_{XW\_ldi\_np}$ by *sil* and get the final transcription S_{sil}. The reason for this is that the *sil* HMM can also model the *context influential* effects of pauses as *ps* and *np* do, and the use of one *context influential* pause HMM can lead to more robust estimations than the use of three HMMs given the same amount of training data. The final acoustic model was trained based on the fixed phonetic transcription S_{sil}:

- *XW_tri_sil3*: Cross-word CD triphone HMMs with the *sil* for both *context influential* pauses and utterance-end silences.

As comparisons, the following cross-word triphone models were also developed.

- *XW_tri_sil1*: Cross-word CD triphone HMMs with the same pause modeling as *XW_tri_sil3*. But the phonetic transcription was achieved by replacing all the *ps* in the $S_{XW\_ldi\_ps}$ by *sil*.
- *XW_tri_sil2*: Cross-word CD triphone HMMs with the same pause modeling as *XW_tri_sil3*. But the phonetic transcription was achieved by replacing all the *ps* in the $S_{XW\_ldi\_dur}$ by *sil*.
- *XW_tri_silon*: Cross-word CD triphone HMMs with the same pause modeling as *XW_tri_sil3*, but the phonetic transcriptions were generated from iterated forced alignments based on evolved HMMs, i.e., the *sil* HMM is initialized from the silences

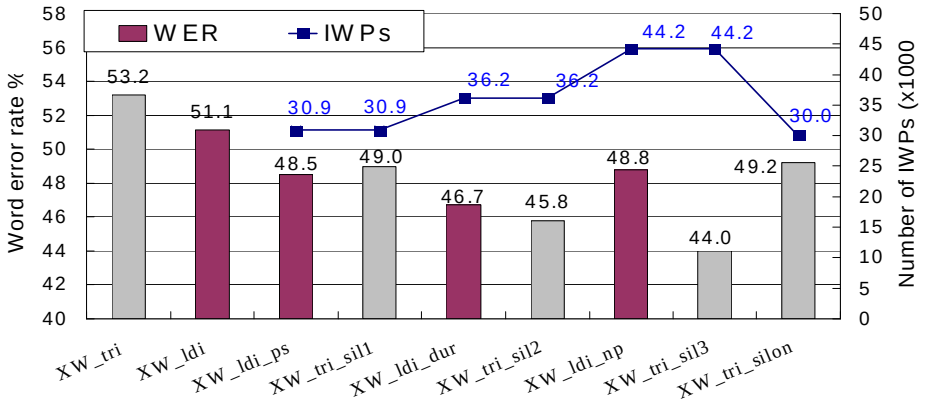

Fig. 3 Recognition results in word error rates for different acoustic models, with the second y-axis denoting the number of detected context influential pauses.

at utterance starts and ends, and it is iteratively trained by assuming it can appear between words with contextual influential effect. This model serves as the baseline system.

5.4 Discussion

Fig. 3 gives the recognition performance for all the acoustic models developed in the previous steps, together with the number of *context influential* pauses detected for training the respective acoustic models. The results suggest:

1. The final triphone HMMs *XW_tri_sil3* achieved the lowest WER among all the acoustic models developed. It got an absolute 5.2% fewer errors than the baseline HMMs *XW_tri_silon*. Since the difference between these two sets of HMMs only lies in the use of different phonetic transcriptions of the training data, the results prove the importance of a correct transcription of pauses on the robustness of acoustic models.
2. Comparison of the three sets of HMMs: *XW_tri_sil1*, *XW_tri_sil2* and *XW_tri_sil3*, clearly shows the positive relation between the number of identified contextual influential pauses and the model's robustness. The increased performance also proved the effectiveness of our proposals to identify pauses and their contextual effects in the approaches of *XW_ldi_ps*, *XW_ldi_dur* and *XW_ldi_np*.
3. The cross-word triphone HMMs *XW_tri* only uses the short pause *sp* HMM to model pauses, and this is the usual way for read speech recognition systems. The absolute 9.2% more errors than the *XW_tri_sil* suggests how significant the effect is of modeling varying pauses for recognizing conversations in noisy environments.

Table 4 gives the environment-based recognition performance for the three representative HMMs developed. Analyses of the results suggest:

Table 4 Noisy environment based Word Error Rates (WER in %) for the three kinds of cross-word triphone HMMs: *XW_tri XW_tri_silon and XW_tri_sil3.*

Env.	SNR	*XW_tri*	*XW_tri_silon*	*XW_tri_sil3*
quiet	23.9 dB	30.0	25.4	23.5
office	20.8 dB	41.0	29.5	28.0
Bradley	15.8 dB	68.6	72.0	57.1
helo	7.6 dB	86.8	84.8	83.3

1. The SNR has a strong correlation to the WERs, with the lowest SNR to the highest WER.
2. For the environments of quiet and stationary office noises, the model *XW_tri_silon* achieved similar improvement to the *XW_tri_sil3* when compared to the *XW_tri* which used *sp* HMM to model only the *Context Free* effect of pauses. This comparison indicates the significance of modeling the *Contextual Influential* effect.
3. The *XW_tri_sil3* got fewer errors than the *XW_tri_silon* under all four conditions, with the most significant 14.9% error reduction for the Bradley noise. It seems that the proposed method showed most effectiveness under this medium level of SNR.
4. Both *XW_tri_silon* and *XW_tri_sil3* only got slight gains for the Helo environment compared to *XW_tri* , indicating that they are not able to efficiently deal with speech of low SNR .

5.5 Pause Statistics

Figure 4 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 the model *XW_tri_sil3*.

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

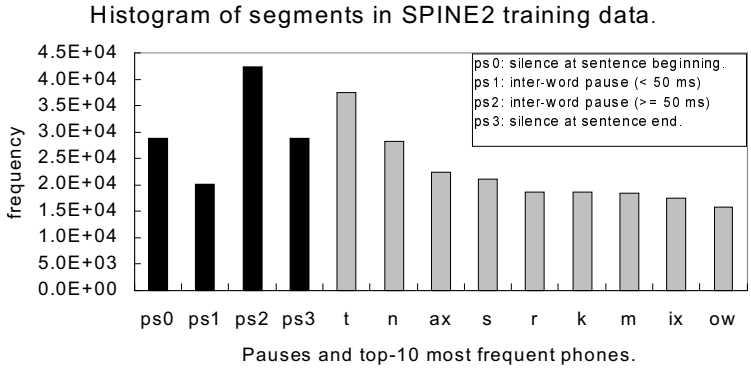

Fig. 4 Histogram for the pauses and top-10 most frequent phones in the training data.

ration is longer than 50ms.

- *ps3* stands for a silence at the end of an utterance.

This shows not only that pauses are very frequent but also that the long inter-word pauses (*ps2*) are the most frequent ones in the data. When they are contaminated by various kinds of background noises, and are not correctly identified in the phonetic segmentation, they will surely hurt the robustness of the developed acoustic models.

6. Conclusions

This paper discussed the influences of varying pauses on building robust acoustic models, and presented an approach to exploit the phonetic heuristics of pauses to achieve more correct phonetic segmentation of the training data. The significantly improved recognition performance proved the efficiency of the proposed approach. The background philosophy is that if acoustic models are accuratedly and robustly developed from the trainning data, then they will have certain robustness against the variations from noises. Since the approach is developed based on only reliable phonetic heuristics, it is suggested to have wide applicability, such as to segmenting speech with background music and etc..

7. ACKNOWLEDGEMENT

We would like to thank Dr. Seiichi Yamamoto for his support for this study, and to acknowledge that the research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, “A study of speech dialogue translation technology based on a large corpus.”

References

[1] B. Zellner, ”Pauses and the temporal structure of speech”, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition, pp. 41-62, John Wiley, Chichester, 1994.

[2] K. Beulen, S. Ortmanns, and Ch. Eliting, ”Dynamic programming search techniques for across-word modeling in speech recognition”, Proc. of IEEE ICASSP’99, pp. 609-612, 1999.

[3] 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.

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

[5] C. Gustafson, and B. Megyesi, ”A comparative study of pauses in dialogues and read speech”, Proc. of Eurospeech 2001, Vol. 2, pp. 931-935, Aalborg, Denmark, 2001.

[6] The second Speech in Noisy Environment Evaluation, http://elazar.itd.nrl.navy.mil/spine/.

[7] T. Kawahara, A. Lee, T. Kobayashi, K. Takeda, N. Minematsu, Sagayama, K. Itoh, A. Itoh, M. Yamamoto, A. Yamada, T.Utsuro, and K. Shikano, ”The Japanese Dictation Software”, http://winnie.kuis.kyoto-u.ac.jp/dictation/, May 8, 2000.

[8] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. Price, ”Segmental durations in the vicinity of prosodic phrase boundaries”, J. of the Acoustical Society of America, 91, pp. 1707-1717, 1992.

[9] D. Byrd, and Ch. Ch. Tan, ”Saying consonant clusters quickly”, J. of Phonetics, 24, pp.263-282, 1996.

[10] D. Byrd, and E. Saltzman, ”Intragestural dynamics of multiple prosodic boundaries”, J. of Phonetics, 26, pp. 173-199, 1998.

[11] J.-S. Zhang, K. Markov, T. Matsui and S. Nakamura, ”Inter-word pauses modeling for recognizing noisy speech in SPINE2 project”, IPSJ SIG Notes, pp. 171-176, Feb., 2002.

[12] N. V. Petlyuchenko, ”Realization of consonant sequences at boundaries of lexical units (instrumental-phonetic research on the material of speech of German radio and television newscasters)”, Ph.D. Dissertation of Odessa I.I. Mechnikov State University, Odessa, Ukraine, 1999.

[13] C. Browman and L. Goldstein, ”Articulatory Phonology: An Overview”, Phonetica 49, pp. 155-180, 1992.

[14] Speech in Noisy Environments (SPINE) Training Audio, http://www.ldc.upenn.edu/Catalog/LDC2000S87.html

[15] 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.

[16] J. Odell, ”The use of context in large vocabulary speech recognition”, Ph. D. Dissertation of the Univ. of Cambridge, March, 1995.

Jin-Song Zhang was born in China. He received his B.E. degree in Electronic Engineering from Hefei University of Technoloy, China, in 1989, an M.E. degree from the University of Science and Technology of China (USTC) in 1992, and a Ph.D. degree from the University of Tokyo, Japan, in 2000. From 1992 to 1996 he worked as a teaching assistant and lecturer in the department of Electronic Engineering of USTC. In 2000, he joined ATR Spoken Language Translation Research Laboratories as an invited researcher. Dr. Zhang is a member of the Acoustic Society of Japan. His main research interests include speech recognition, prosody information processing, and speech synthesis.

Konstantin Markov was born in Bulgaria. He received his B.E. degree in Electrical Engineering from Department of Cybernetics, St. Petersburg Technical University, Russia in 1984. In 1996 and 1999 he received his M.E. and D.E. in Computer Science from Toyohashi University of Technology, Japan. From 1999 to 2000, he was a research engineer in ATR Research and development center and from 2000 he joined ATR Spoken Language Translation Research Laboratories as invited researcher. Dr. Markov is member of Acoustic Society of Japan. His main research interests include automatic speech recognition noise robustness, speaker identification and statistical pattern recognition in general.

Tomoko Matsui received the Ph. D. degree in computer science from Tokyo Institute of Technology, Tokyo, in 1997. She has been a permanent researcher at NTT since 1988. From January to June in 2001, she was a member of the Acoustic and Speech Research Department, Bell Laboratories, Murray Hill, NJ, as a visiting researcher working on confidence measure for speech recognition. She is currently visiting ATR, Kyoto, as a senior researcher working on speech recognition. Her research interests include speech and speaker recognition. She received the paper award from the Institute of Electronics, Information and Communication Engineers of Japan (IEICE) (1993).

Satoshi Nakamura was born in Japan on August 4, 1958. He received the B.S. degree in electronics engineering from Kyoto Institute of Technology in 1981 and the Ph.D. degree in information science from Kyoto University in 1992. Between 1981-1986 and 1990-1993, he worked with the Central Research Laboratory, Sharp Corporation, Nara, Japan, where he was engaged in speech recognition research. During 1986-1989, he was@a researcher of the speech processing department at ATR Interpreting Telephony Research Laboratories. From 1994-2000, he was an associate professor of the graduate school of information science, Nara Institute of Science and Technology, Japan. In 1996, he was a visiting research professor of the CAIP center of Rutgers, the state university of New Jersey USA. He is currently the head of Department 1 in ATR Spoken Language Translation Laboratories, Japan. He also serves as a guest professor for Toyohashi University of Technology and Ritsumeikan University from April, 2002. His current research interests include speech recognition, speech translation, spoken dialogue systems, stochastic modeling of speech, and microphone arrays. He received the Awaya Award from the Acoustical Society of Japan in 1992, and the Interaction2001 best paper award from the Information Processing Society of Japan in 2001. He is a member of the Acoustical Society of Japan, Institute of Electrical and Electronics Engineers (IEICE), Information Processing Society of Japan, and IEEE. He is currently a member of the Speech Technical Committee of the IEEE Signal Processing Society and an editor for the Journal of the IEICE Information and System Society.