MODELING VARYING PAUSES TO DEVELOP ROBUST ACOUSTIC MODELS FOR RECOGNIZING NOISY CONVERSATIONAL SPEECH

Jin-Song Zhang and Satoshi Nakamura
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-0288 Japan
{jinsong.zhang, satoshi.nakamura}@atr.co.jp

ABSTRACT
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 our proposal to exploit 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 SPINEII project, and achieved a more correct phonetic transcription. The cross-word triphone HMMs developed using this transcription got absolute 5.2% word error reduction when compared to the baseline model.

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 existences of silences and intra-segmental pauses are relatively more stable than those 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 the recognition performances. 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 by about 5% relative error reductions than the one 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 over than 1% absolute error reductions 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. As 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 further more miss-segment those varying pauses in a later iteration of forced alignment. This kind of circles 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]. 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 done in the conventional approaches, we took the segmentation for 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 pauses’ HMMs. Such initialized HMMs lead to better final HMMs and phonetic segmentations. After the optimized transcription became available, it can be used to train a set of more robust acoustic models.

Studies have been made on the data of “Speech in Noisy Environments II” (SPINEII) project, 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 pauses’ HMMs and the phonetic transcriptions. Experimental results showed that the approach effectively detected increasing number of noisy pauses, and the final cross-word triphone HMMs based on the optimized phonetic transcription achieved by 5.2% less word errors than the baseline HMMs.

The following is arranged as: Section 2 describes the precise contextual modeling of pauses, and explains why a correct phonetic transcription is important. Section 3 introduces the proposal to exploit phonetic heuristics about pauses for collecting initialization samples for pauses’ HMMs. Section 4 introduces the SPINEII data and experimental set-up. Section 5 presents the experimental results and discussions. Finally, section 6 gives conclusions.

This research was supported in part by the Telecommunications Advancement Organization of Japan.
2. PRECISE CONTEXTUAL MODELING OF PAUSES

In order to achieve robust cross-word context dependent HMMs, we adopt precise contextual modeling of pauses, which uses several pause HMMs with different contextual effects to model the specific coarticulations of the phone context of a pause. In this study the following two kinds of contextual effects are used:

- **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 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 it. Such a pause appears in the context of its neighboring triphones to indicate its context influential effects. When use the "sil" HMM for these pauses [4], the triphones for the "B pause C" in the above example would become as "A-B+sil sil sil-C+D".

Thus, for the same pause segment in one utterance, a different identification of its contextual effects will result in using not only a different pause HMM but also two different neighboring allophone HMMs in the phonetic transcription. If the pauses and their contextual effects are frequently miss-segmented and miss-identified for the training data, then both the pause HMMs and their neighboring allophone HMMs will be badly estimated. Therefore, a correct phonetic segmentation is important for achieving a set of robust acoustic models.

3. PHONETIC HEURISTICS ORIGINATED PAUSE MODELING

The failure to use forced alignments to segment varying pauses in noisy speech can be ascribed to the defect of the bootstrapping way to develop HMMs, which desires robustly initialized HMMs. If we can get enough samples of noisy pauses for initializing the pauses’ HMMs, the bootstrapping way will lead to better HMMs and phonetic segmentations. For this purpose, we propose to exploit reliable phonetic heuristics of pauses to find sufficient initialization samples for pause HMMs.

The first important characteristic of pauses related to acoustic modeling is the relations between the length of a pause and the phone coarticulations across it [6, 7]. 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, [7] showed that a pause longer than 60ms promotes preservation of distinctive features of consonants at the boundary. This suggests that 50-70ms be 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 detections is their important relations to the prosodic phrasing structure of speech, which has syntactic and semantic originations[1, 5]. Speakers tend to pause more frequently and longer at word boundaries which are associated with higher levels in the prosodic hierarchy. Besides, 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 influences on the following phones. This means that a pause would probably appear after or before an extra-ordinarily long phone at a word boundary.

Based on the two heuristics, we propose the following method to generate initialization samples for pauses.

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 the others shorter than 50ms as context free.

3. **Phrase boundary pauses insertion:** extra context influential pauses can be inserted to the transcription S at places where prosody phrase boundaries are assumed existing, based on statistical phone duration analysis. The method is:

   - First, compute the duration mean \( \mu_i \) and deviation \( \sigma_i \) 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 \( (> \mu_i + (2\sim3) \times \sigma_i) \), a context influential pause label would be inserted for a possibly miss-located pause.

4. SPINEII DATA AND EXPERIMENTAL SET-UP

The SPINEII data is organized in conversations between two speakers collaborating in a task of seeking and shooting targets, with reproduced previously recorded background noises. The 11 types of noisy environments, including quiet, office, aircraft carrier, street, car, helicopter, tank, fighter jet and others, 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 [8]. 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.

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 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 5 dB to 20 dB [8]. All data
have only transcripts at word level, no phonetic segment-
ation information.

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

4.2. Experimental Set-up
The acoustic feature used in this study is the standard mel-
scale cepstrum (MFCC), which is computed with a frame
size of 20ms and frame shift of 10 ms. 12 MFCCs plus log
energy and their 1st and 2nd order time derivatives form a
39 dimensional vector.

The basic lexicon consists of about 5,7k unique words
with a total of about 11k entries for pronunciation variation-
s. The basic phone set has 43 American phones, a sp for
short pauses and a sil for pauses and silences at utterance-
ends. All the phone HMMs have 3 left-to-right states, ex-
cept that the sp has only one skipable 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.

5. EXPERIMENTS
Due to the rather complex variations in the SPINE data,
a stepwise procedure based on the phonetic heuristics o-
riginated pause modeling was used to optimize the pauses’
HMMs and phonetic segmentation, after a preliminary in-
vestigation 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 hav-
ing about 2,000 phonetic decision tree based tied states and
16 mixtures of Gaussians per state.

- \textit{XW}_{\text{tri}}: cross-word CD tri-phones.
- \textit{IW}_{\text{tri}}: intra-word CD tri-phones.
- \textit{XW}_{\text{ldi}}: cross-word left CD di-phones.
- \textit{XW}_{\text{rdi}}: cross-word right CD di-phones.

The four models used the sp for pauses and the sil for
silences at utterance-ends. Fig. 1 illustrates the recogni-
tion performances in word error rates (WER), showing
that \textit{XW}_{\text{ldi}}, the cross-word left CD diphone HMMs, is the
most robust one in this case. The probable reason may be
attributed to the factor of different average training sam-
ple per allophone (ASP) of each modeling: the \textit{XW}_{\text{ldi}} has
the highest ASP value which leads to robustness against
the influences from incorrect pauses’ modeling. (The bet-
ter performance of \textit{XW}_{\text{ldi}} than \textit{XW}_{\text{rdi}} may reflect the fact
that carryover coarticulations are more significant than the
anticipations in natural speech.) Therefore, cross-word left

![Fig. 1. Word error rates (WER) and average training samples per allophone (ASP) of different context modeling.](image)

CD diphone modeling was chosen to be used in the lat-
er steps to optimize phonetic segmentations. The model
\textit{XW}_{\text{ldi}} generated a phonetic transcription \textit{SXW}_{\text{ldi}}.

5.2. Stepwise Optimization of Pause Modeling and
Phonetic Segmentation
- **Step 1:** Incorporate a \textit{ps} HMM for pauses to model
the phenomenon that pauses may be varying from si-
lences. The \textit{ps} HMM is context \textit{influential}, and its ini-
tialization samples took those \textit{sp} segments in \textit{SXW}_{\text{ldi}}
whose durations were longer than 50ms. Through
the iterations of model estimations and forced align-
ments, we developed a new set of cross-word left CD
diphone HMMs \textit{XW}_{\text{ldi}}\textit{ps} and got a new phonetic
transcriptions \textit{SXW}_{\text{ldi}}\textit{ps}.

- **Step 2:** Insert \textit{ps} labels into \textit{SXW}_{\text{ldi}}\textit{ps}, according to
the method of \textit{phrase boundary pauses insertion} in or-
der to generate initialization samples for those pauses
with high-level noises. Then, through the same iter-
ations as Step 1, we got new HMMs \textit{XW}_{\text{ldi}}\textit{ps} and
transcription \textit{SXW}_{\text{ldi}}\textit{ps}.

- **Step 3:** Incorporate another \textit{np} HMM as more de-
tailed noisy pauses’ model. The \textit{np} HMM is also
context \textit{influential}, and its initialization samples took
those \textit{ps} in utterances from noisy channels based on
\textit{SXW}_{\text{ldi}}\textit{ps}.

In each step, speech recognition experiments were car-
rried out based on the developed HMMs in order to show the
efficiencies.

5.3. Developing Final Acoustic Models
The phonetic transcription \textit{SXW}_{\text{ldi}}\textit{ps} is assumed contain-
ing rather correct segmentation of pauses. Then we replace
all the labels of \textit{ps} and \textit{np} in \textit{SXW}_{\text{ldi}}\textit{ps} by \textit{sil} and get the
final transcription \textit{SXW}_{\text{sil}}. The reason is that the \textit{sil} HMM

- **Step 1:** Incorporate a \textit{ps} HMM for pauses to model
the phenomenon that pauses may be varying from si-
lences. The \textit{ps} HMM is context \textit{influential}, and its ini-
tialization samples took those \textit{sp} segments in \textit{SXW}_{\text{ldi}}
whose durations were longer than 50ms. Through
the iterations of model estimations and forced align-
ments, we developed a new set of cross-word left CD
diphone HMMs \textit{XW}_{\text{ldi}}\textit{ps} and got a new phonetic
transcriptions \textit{SXW}_{\text{ldi}}\textit{ps}.

- **Step 2:** Insert \textit{ps} labels into \textit{SXW}_{\text{ldi}}\textit{ps}, according to
the method of \textit{phrase boundary pauses insertion} in or-
der to generate initialization samples for those pauses
with high-level noises. Then, through the same iter-
ations as Step 1, we got new HMMs \textit{XW}_{\text{ldi}}\textit{ps} and
transcription \textit{SXW}_{\text{ldi}}\textit{ps}.

In each step, speech recognition experiments were car-
rried out based on the developed HMMs in order to show the
efficiencies.

5.3. Developing Final Acoustic Models
The phonetic transcription \textit{SXW}_{\text{ldi}}\textit{ps} is assumed contain-
ing rather correct segmentation of pauses. Then we replace
all the labels of \textit{ps} and \textit{np} in \textit{SXW}_{\text{ldi}}\textit{ps} by \textit{sil} and get the
final transcription \textit{SXW}_{\text{sil}}. The reason is that the \textit{sil} HMM

- **Step 1:** Incorporate a \textit{ps} HMM for pauses to model
the phenomenon that pauses may be varying from si-
lences. The \textit{ps} HMM is context \textit{influential}, and its ini-
tialization samples took those \textit{sp} segments in \textit{SXW}_{\text{ldi}}
whose durations were longer than 50ms. Through
the iterations of model estimations and forced align-
ments, we developed a new set of cross-word left CD
diphone HMMs \textit{XW}_{\text{ldi}}\textit{ps} and got a new phonetic
transcriptions \textit{SXW}_{\text{ldi}}\textit{ps}.

In each step, speech recognition experiments were car-
rried out based on the developed HMMs in order to show the
efficiencies.

5.3. Developing Final Acoustic Models
The phonetic transcription \textit{SXW}_{\text{ldi}}\textit{ps} is assumed contain-
ing rather correct segmentation of pauses. Then we replace
all the labels of \textit{ps} and \textit{np} in \textit{SXW}_{\text{ldi}}\textit{ps} by \textit{sil} and get the
final transcription \textit{SXW}_{\text{sil}}. The reason is that the \textit{sil} HMM

- **Step 1:** Incorporate a \textit{ps} HMM for pauses to model
the phenomenon that pauses may be varying from si-
lences. The \textit{ps} HMM is context \textit{influential}, and its ini-
tialization samples took those \textit{sp} segments in \textit{SXW}_{\text{ldi}}
whose durations were longer than 50ms. Through
the iterations of model estimations and forced align-
ments, we developed a new set of cross-word left CD
diphone HMMs \textit{XW}_{\text{ldi}}\textit{ps} and got a new phonetic
transcriptions \textit{SXW}_{\text{ldi}}\textit{ps}.

In each step, speech recognition experiments were car-
rried out based on the developed HMMs in order to show the
efficiencies.

5.3. Developing Final Acoustic Models
The phonetic transcription \textit{SXW}_{\text{ldi}}\textit{ps} is assumed contain-
ing rather correct segmentation of pauses. Then we replace
all the labels of \textit{ps} and \textit{np} in \textit{SXW}_{\text{ldi}}\textit{ps} by \textit{sil} and get the
final transcription \textit{SXW}_{\text{sil}}. The reason is that the \textit{sil} HMM

- **Step 1:** Incorporate a \textit{ps} HMM for pauses to model
the phenomenon that pauses may be varying from si-
lences. The \textit{ps} HMM is context \textit{influential}, and its ini-
tialization samples took those \textit{sp} segments in \textit{SXW}_{\text{ldi}}
whose durations were longer than 50ms. Through
the iterations of model estimations and forced align-
ments, we developed a new set of cross-word left CD
diphone HMMs \textit{XW}_{\text{ldi}}\textit{ps} and got a new phonetic
transcriptions \textit{SXW}_{\text{ldi}}\textit{ps}.

In each step, speech recognition experiments were car-
rried out based on the developed HMMs in order to show the
efficiencies.
- $X_{W_{tri\_sil}}$: Cross-word CD triphone HMMs with the sil for both context influential pauses and utterance-end silences.

As a comparison, the following model was developed as the baseline:
- $X_{W_{tri\_sil/cilon}}$: Cross-word CD triphone HMMs with the same pause modeling as $X_{W_{tri\_sil}}$, but the phonetic transcriptions were generated from iterated forced alignments based on evolved HMMs.

![Recognition results](image)

**Fig. 2.** Recognition results in word error rates for different acoustic models, with the second y-axis denoting the number of detected context influential pauses.

### 5.4. Discussions:

Fig. 2 gives the recognition performances 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 cross-word CD triphone HMMs $X_{W_{tri\_sil}}$ achieved the lowest WER among all the acoustic models developed. It got absolute 5.2% less errors than the baseline HMMs $X_{W_{tri\_sil/cilon}}$. Since the difference between these two models 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, and the efficiency of the optimization approaches for pause modeling and phonetic segmentation.

2. The cross-word triphone HMMs $X_{W_{tri}}$ only use 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 $X_{W_{tri\_sil/cilon}}$ suggest how significant the effect is of modeling varying pauses for recognizing conversations in noisy environments.

3. When looking at the models $X_{W_{tri\_lps}}$, $X_{W_{tri\_lps\_dur}}$, and $X_{W_{tri\_lps\_dur}}$, the gradually reduced WERs and the increasing number of detected pauses showed the effectiveness of the approaches, including the incorporation of the $ps$ HMM and the proposed methods of pause initialization based on phonetic heuristics.

4. The model $X_{W_{tri\_lps\_np}}$ incorporating the $np$ HMM effectively detected more 8k tokens of pauses than the $X_{W_{tri\_lps\_dur}}$. The increased WER may be ascribed to the biased estimations of pause-related allophone HMMs to the training data.

5. We also want to point out that the WERs here are significantly higher than those by other groups such as CMU [8], as they are not comparable. One reason is due to different size of testing sets, and the second reason is that neither multi-session adaptations nor decoding optimizations were used here. Exactly, only CMN and optimization of language model scale brought about by absolute 11% error reductions to the model $X_{W_{tri\_sil/cilon}}$.

### 6. CONCLUSIONS

This paper discussed the influences of varying pauses on building robust acoustic models, and presented our approach to exploit phonetic heuristics of pauses to achieve more correct phonetic segmentation of the training data. The significantly improved recognition performances proved the efficiency of the proposed approaches. The authors suggest the proposed methods have general applicability.

**ACKNOWLEDGEMENT:** We would like to thank Dr. Konstantin Markov, Dr. Tomoko Matsui, Mr. Rainier Gruhn, and other colleagues for their help and valuable discussions during the study.

### 7. REFERENCES


