PAUSE AS A PHRASE DEMARCATOR FOR SPEECH AND LANGUAGE PROCESSING

JUNKO HOSAKA MARK SELIGMAN HARALD SINGER

ATR Interpreting Telephony Research Laboratories
Hikaridai 2-2, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Abstract

In spontaneous speech understanding a sophisticated integration of speech recognition and language processing is especially crucial. However, the two modules are traditionally designed independently, with independent linguistic rules. In Japanese speech recognition the *bunsetsu* phrase is the basic processing unit and in language processing the sentence is the basic unit. This difference has made it impractical to use a unique set of linguistic rules for both types of processing. Further, spontaneous speech contains unexpected utterances other than well-formed sentences, while linguistic rules for both speech and language processing expect well-formed sentences. They therefore fail to process everyday spoken language. To bridge the gap between speech and language processing, we propose that pauses be treated as phrase demarcators and that the interpausal phrase be the basic common processing unit. And to treat the linguistic phenomena of spoken language properly, we survey relevant features in spontaneous speech data. We then examine the effect of integrating pausal and spontaneous speech phenomena into syntactic rules for speech recognition, using 118 sentences. Our experiments show that incorporating pausal phenomena as purely syntactic constraints degrades recognition accuracy considerably, while the additional degradation is minor if some further spontaneous speech features are also incorporated.

1 INTRODUCTION

A spontaneous speech understanding system accepts naturally spoken input and understands its meaning. In such a system, speech processing and language processing must be integrated in a sophisticated manner. However, the integration is not straightforward, as the two are studied independently and have different processing units. Moreover, spontaneous speech contains unexpected phenomena, such as hesitations, corrections and fragmentary expressions, which thus far have not been treated in linguistic rules.

The most significant concern in speech processing is raising the recognition accuracy. For that purpose, applying linguistic information, e.g. using stochastic models[1], syntactic rules[2], semantic information[3] and discourse plans[4], is most promising. In a recent Japanese speech translation system[5] *bunsetsu*-based syntactic constraints are successfully applied in the speech processing module[6][1]. However, rules representing the same constraints cannot be used directly in sentence-based language processing, where the primary concern is to understand sentence meaning. In speech recognition, a sequence of words forms a *bunsetsu* and a set of *bunsetsus* then forms a sentence. In language processing, on the other hand, where the sentence is the basic processing unit, treating the main verb and its complements is usually the core of processing. For the sentence *kaigi ni moshikomi tai no desu ga*, meaning 'I would like to apply for the conference,' the processing discrepancy is sketched in Figure 1:

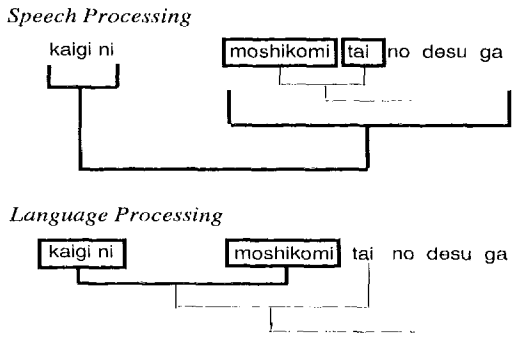

Figure 1: Structural Difference

Although linguistic rules for speech recognition always cope with uncertain phoneme hypotheses, they still expect well-formed speech input, and this is even more true of linguistic rules in language processing. In spontaneous speech, however, there are hesitations, corrections and incomplete utterances which are not treated in the conventional framework.

In addressing spontaneous speech understanding, two main problems must be solved: the absence of common processing components as sketched in Figure 1, and our insufficient knowledge of spontaneous speech features. In this paper, we propose the pause as a phrase demarcator and the interpausal phrase as the basic processing unit. A phrase is naturally demarcated with pauses in spoken language and an interpausal phrase often functions as a meaning unit[8][9]. In spontaneous speech understanding we must both accept naturally spoken input and understand its meaning. Use of the pause as a phrase demarcator is advantageous for both of these purposes. Further, we investigate several frequent spontaneous

[1] A *bunsetsu* roughly corresponds to a phrase and is the next largest unit after the word. The number of words in a phrase ranges from 1 to 14, and the mean number is about 3[7].

speech features using spontaneous speech data[10]. We then apply the study to speech recognition. We examine the effect of integrating into syntactic rules pausal phenomena and certain features of spoken language, using 118 test sentences.

2 ANALYSIS OF SPONTANEOUS DIALOGUES

2.1 Spontaneous Dialogue Data

As sources of spontaneous data, we use four Japanese dialogues concerning directions from Kyoto station to either a conference center or a hotel, collected in the Environment for Multi-Modal Interaction[10]. Speaker A is pre-trained to give the directions, mentioning possible transportation, location and so forth. Two subjects seeking directions, Speaker B and Speaker C, are given some keywords, such as the name and the date of the conference. They may use telephone connections only, or may use a multimodal setup with onscreen graphics and video as well. Table 1 shows how many words are used in the dialogues studied:

Table 1: Words in the Corpora

	Telephone	Multimedia
Speakers A,B	536	714
Speakers A,C	1167	1124
Subtotal	1703	1838
Total	3541	

The corpora consists of 3541 words in total, and contains 440 different words. It has 403 turn-takings, and thus roughly 403 sentences.

In the multimedia setup, speakers use deictic expressions such as *koko* and *kore* meaning "here" and "this," respectively. The dialogues also lasted longer than those in the telephone-only setup. However, we did not find any further distinct differences between the two setups. We therefore analyse all of the dialogues in the same way.

For our study, transcripts of the spontaneous dialogues have been prepared, and these contain morphological tags and turn-taking information. Pause information within turns, i.e., breaths or silences longer than 400 miliseconds, is provided as well.

2.2 Pause as a Phrase Demarcator

In Table 2 we illustrate the adequacy of the interpausal phrase as a processing unit with a series of directions to Kyoto station's Karasumachou exit. The entire explanation consists of three turns separated by short response syllables, such as *hai*, that do not overlap the explanation. That is, the speaker paused during these responses. We marked each turn with TURN at the end. As a primary demarcator we used pauses and turns. Thus either PAUSE or TURN appears in the second column. Further demarcator candidates such as the filled pauses *anoo* or *eeto*, the emphasis marker *desune* and the response syllable *hai* when overlapping the explanation appear in the third column as FILLED PAUSE, DESUNE and RESPONSE, respectively. A rough translation follows each interpausal phrase:

Table 2: Phrase Demarcator

ここからでしたら *if it is from here*	PAUSE	FILLED PAUSE
こちら *this side*	PAUSE	
の階段を上がって頂きまして *you go up the stairs*	PAUSE	RESPONSE
ここずっと渡って頂きます *you cross here all the way*	TURN	
で *and*	PAUSE	
次の階段こちらが見えてきた 時点で 左に曲がって頂きます 一番最初に *when you see the next stairs, this one, turn left, first*	PAUSE	RESPONSE
出てきます この交差点みたいな ところを *at this place like a crossroad which appears*	PAUSE	DESUNE
右に行って頂きます *turn right*	TURN	
で右に行って頂いて *and you turn right*	PAUSE	
でここの階段を 降りて頂きますと *and then if you go down the stairs here*	PAUSE	RESPONSE
烏丸町口に出てまいります *you come out of the karasumachou exit*	TURN	

The length of the processing unit plays an important role in speech recognition. Table 2 shows that alternative demarcator candidates such as FILLED PAUSE and RESPONSE usually cooccur with pauses. In Table 2, for example, we find only one case where RESPONSE does not cooccur with a pause. Consequently, the segments within turns bounded by these alternative markers would not be much different from those bounded by pauses; in particular, they would not be much shorter or longer. Thus, at least where length is concerned, the combination of PAUSE and TURN seems appropriate and sufficient to mark out phrases. With respect to language processing, Table 2 shows that interpausal phrases are often adequate as translation units, which suggests that such phrases often function as meaning units.

Interpausal phrases typically end with a conjunctive postposition, such as *ga* or *keredomo*; a postpositional phrase; an interjection, such as *hai* or *moshimoshi*; the genitive postposition *no* for adnominals;

an adnominal conjugation form; a coordinate conjugation form; auxiliaries with sentence final conjugation form; or a sentence final particle, such as *ka* or *ne*.

2.3 Features of Spontaneous Dialogues

We studied ten features of spontaneous dialogues which are not considered in grammars for well-formed sentences[6][11]. Table 3 shows the features and their frequencies:

Table 3: Feature and Occurrence

Use of *desune*	37
Use of *anoo*	35
Fragmentary utterance	26
Use of *ecto*	15
End of turn with a PP	7
Postposition drop	7
Question without *ka*	5
Disfluency: *soudesune*	5
Apposition	4
Inversion	3

We expected a very high frequency of the filled pauses *anoo* and *ecto* functioning as discourse managers[12]. However, Table 3 shows only a modest frequency. Phonological variations such as *attnoo* and *ano* for *anoo* and *ettoo* and *ectto* for *ecto* were not counted. This may be why the frequency of both expressions is unexpectedly low.

Some features shown in Table 3 are discussed in the example sets below. Features in focus are in bold type:

Ex. 1
sochira no **desune** *noriba kara basu ga* **desune** *detemasu*
there is a bus from that bus stop

The person giving directions often uses the expression *desune*. The use of *desune* emphasizes the preceding utterance, typically the immediately preceding minimal phrase. In Ex. 1 the first use emphasizes *sochira no* and the second stresses *basu ga*.

We denote the person giving the directions as Speaker A and the person seeking the information as Speaker B in Examples 2, and 3.

Ex. 2

Speaker B:	*keage no* **kita**
	north of keage
Speaker A:	*sou desu*
	that's right
Speaker B:	**deguchi**
	exit
Speaker A:	*kitadeguchi desu ne*
	it's the north exit, okay?

In Ex. 2 Speaker B did not finish what he wanted to say, but Speaker A understood his intention and interrupted his utterance, which is therefore fragmentary. Speaker B continued but before he could finish Speaker A finished for him. So Speaker B's utterance is again fragmentary.

Ex. 3

Speaker A:	**futaeki de**
	after two stops
Speaker B:	*keage*
	keage
Speaker A:	*sou desu*
	that's right

Speaker A is giving directions but before he has completed his utterance Speaker B interrupts with the station name. Speaker A did not continue his first utterance and agreed with Speaker B. Speaker A's first utterance is a nominal phrase, which is never completed.

3 APPLICATION OF THE ANALYSIS

To examine the feasibility of integrating into syntactic rules both pausal phenomena and the features of spontaneous speech studied in Section 2, we prepared three different sets of rules. In all three sets, rules have been explicitly modified to represent pausal phenomena. The first set `Pause` contains only such modifications, while the other two sets add one additional spontaneous feature each: rule set `Emphasis` permits use of the emphasis marker *desune* after a noun phrase, while rule set `Turn` allows postpositional utterances at the end of a turn. We conducted preliminary speech recognition experiments with a parser which uses linguistic constraints written as a CFG.

3.1 Linguistic Constraints

To represent our underlying linguistic constraints we adapted existing syntactic rules developed for speech recognition[6]. Earlier experiments using *bunsetsu*-based speech input showed 70% sentence recognition accuracy for the top candidate and 84% for the top 5 candidates.

The format for all of our syntactic rules is as follows:

```
(<CAT1> <--> (<CAT2> <CAT3>))
```

Nonterminals are surrounded by <>. The above rule indicates that `CAT1` consists of `CAT2` and `CAT3`. We denote the categories in interpausal phrase rules in lower-case and the categories in interpausal phrase-based sentence rules in upper-case.

In the rule set `Pause` we prepared about 45 phrases that can end with a pause: postpositional phrases, conjunctive phrases, adnominal verbal phrases marked with a special conjugation form,

phrases that end with a conjunctive postposition, adnominal phrases with the genitive postposition *no*, and coordinate verbal phrases. The first three rules are as follows:

```
(<pp-pau> <--> (<pp> <pause>))
(<conj-pau> <--> (<conj> <pause>))
(<vaux-mod-pau> <--> (<vaux-mod> <pause>))
```

In the rule set `Emphasis` we prepared seven additional rules for treating the emphasis marker *desune*, represented as follows:

```
(<pp-pau> <--> (<pp> <emphasis> <pause>))
(<pp-no-pau> <-->
            (<pp-no> <emphasis> <pause>))
```

Methods for combining interpausal phrases to obtain an overall utterance meaning require further study. At this stage we defined a sentence very loosely. It can be an interjection; an interjection followed by a combination of interpausal phrases; or simply a combination of interpausal phrases. To allow fragmentary utterances, in the rule set `Turn`, we also introduced a sentence consisting of a nominal phrase, which may contain adnominal phrases. Complete sentences in `Turn` are defined as follows:

```
(<SSS> <--> (<INTERJ1>))
(<SSS> <--> (<INTERJ1> <SS>))
(<SSS> <--> (<SS>))
(<SSS> <--> (<M-NN>))
```

Table 4 shows the size and phoneme perplexity of the three sets of rules:

Table 4: Size and Perplexity

	Pause	Emphasis	Turn
Rules	2326	2333	2327
Words	751	752	751
Perplexity	3.96	3.96	3.96

A given phoneme string can belong to several categories. For instance, *de* can be a postposition or a copula conjugation form. The number of different phoneme strings is 503 for `Pause` and `Turn`, and 504 for `Emphasis`.

3.2 Speech Recognition Experiment

We conducted a speech recognition experiment with 118 test sentences concerning secretarial services for an international conference. A professional broadcaster uttered the sentences without any special constraints such as pause placement.

For our speech recognition parser, we used HMM-LR[14], which is a combination of generalized LR parsing and Hidden Markov Models (HMM). The system predicts phonemes by using an LR parsing table and drives HMM phoneme verifiers to detect or verify them without any intervening structure such as a phoneme lattice. Linguistic rules for parsing can be written in CFG format.

As mentioned in section 3.1, we explicitly defined rules that can end with pauses in linguistic constraints. According to the pause model, a pause can last from 1 to 150 frames, where a frame lasts 9 msec.

Examples (1) and (2) show the results of HMM-LR Japanese speech recognition[2]. (1) shows sample results of rule set `Pause` and (2) shows sample results of `Turn`. The phoneme strings which were actually pronounced are enclosed in | |:

```
(1)  |kaiginoaNnaishowaomochidesuka|
     (Do you have a conference invitation?)
-----------------------------------------------
  1: kaigi-no-P-aNnaisyo-o-omochi-desu-ka
  2: kaigi-ni-P-aNnaisyo-o-omochi-desu-ka
  3: kaigi-ga-P-aNnaisyo-o-omochi-desu-ka
> 4: kaigi-no-P-aNnaisyo-wa-P-omochi-desu-ka
  5: kaigi-ni-P-aNnaisyo-wa-P-omochi-desu-ka

(2) |iie|(no)
-----------------------------------------------
  1: imi-e
  2: igo-e
> 3: iie
  4: ima-e
  5: kigeN-e
```

In the examples, the symbols >, -, `N` and `P` have special meaning: A correctly recognized phrase is marked with >. A word boundary is marked with -. A syllabic nasal is transcribed `N`. A pause is marked with `P`.

Example (1) shows typical recognition errors involving postpositions like *no, ni, ga,* and *o,* which often receive reduced pronunciation in natural speech. The surounding context may aggravate the problem. Here, for instance, topic marker *wa* is erroneously recognized as object marker *o* in the environment of preceding and subsequent phoneme *o*. The possible introduction of pauses at such junctures further complicates the recognition problem. Analysis deeper than CFG parsing will often be needed to filter unlikely candidates. Example (2) demonstrates the dangers of allowing postpositional phrases to end utterances. Here, all recognition candidates other than the third are inappropriate postpositional phrases. To recognize the unlikelihood of such candidates, we will need further controls, such as discourse management.

Our resulting sentence speech recognition accuracies are shown in Table 5. For instance, using rule set `Pause`, the correct candidate was the highest ranking candidate 50.0 percent of the time, Rank 1, while the correct candidate was among the top 5 candidates 55.9 percent of the time, Rank 5.

[2]The maximal amount of the whole beam width, called the global beam width, is set at 100, and the maximal beam width of each branch, the local beam width, is 12.

Table 5: Recognition Rate (%)

Rank	Pause	Emphasis	Turn
1	50.0	50.0	46.6
2	54.2	54.2	53.4
3	55.1	55.1	55.1
4	55.9	55.9	55.9
5	55.9	55.9	55.9

With the underlying linguistic rules for the three rule sets, earlier experiments had achieved 70% sentence speech recognition accuracy for speech input with explicit pauses at *bunsetsu* boundaries. Our best present results for spontaneous speech are much more modest: 50%.

Table 5 shows that the introduction of the emphasis marker *desune* did not affect processing: as seen in Table 4, rule set `Emphasis` has a slightly higher perplexity than `Pause`, but we had exactly the same results for the two. On the other hand, the perplexities of `Pause` and `Turn` are identical, but the treatment of fragmentary utterances did decrease recognition accuracy.

4 CONCLUSION

To treat spontaneous speech understanding we have two main problems: the absence of a common processing unit and insufficient knowledge of spontaneous speech features.

We have proposed pauses as phrase demarcators and interpausal phrases as common processing units to allow integration of speech recognition and language processing in the processing of spontaneous speech understanding. We demonstrated the advantages of processing based on interpausal phrases using examples taken from spontaneous speech dialogues containing 3541 words. Using the same data, we studied certain features of spoken language, such as filled pauses and fragmentary utterances. Based on the study, we prepared three different CFG rule sets for preliminary speech recognition experiments. In all three sets, rules have been explicitly modified to represent pausal phenomena. The first set contains only such modifications, while the other two sets add one additional spontaneous feature each: use of the emphasis marker *desune* after a noun phrase or postpositional utterances at the end of a turn. For 118 sentences, sentence recognition accuracy for pause-based rules was considerably less than the accuracy obtained in earlier *bunsetsu*-based tests using mandatory pauses at *bunsetsu* boundaries; but further loss of accuracy caused by incorporating the spontaneous features was minor.

We believe that the loss of speech recognition accuracy for sentences seen in our pause-based experiments is largely due to the difficulties of combining interpausal phrase hypotheses. Our rules currently combine interpausal phrases in a relatively unconstrained manner, using only weak syntactic constraints. Based on further study of the structures which precede and follow pauses or filled pauses, we hope to provide stronger syntactic constraints in the future.

5 ACKNOWLEDGEMENTS

We wish to thank Dr. Y. Yamazaki, President of ATR-ITL, T. Morimoto, Head of Department 4, and many of our ITL colleagues for their generous support and encouragement.

References

[1] Lee, K.-F. and Hon, H.-W.(1988): "Large-Vocabulary Speaker-Independent Continuous Speech Recognition Using HMM," Proc. of ICASSP-88, pp.123-126.

[2] Ney, H.(1987): "Dynamic Programming Speech Recognition Using a Context-Free Grammar," Proc. of ICASSP-87, pp.69-72.

[3] Matsunaga, S., Sagayama, S., Homma, S. and Furui, S.(1990): "A Continuous Speech Recognition System Based on a Two-Level Grammar Approach," Proc. of ICASSP-90, pp.589-592.

[4] Yamaoka, T. and Iida, H.(1990): "A Method to Predict the Next Utterance Using a Four-layered Plan Recognition Model," Proc. of ECAI-90, pp.726-731.

[5] Morimoto, T., Takezawa, T., Yato, F., et al.(1993): "ATR's Speech Translation System: ASURA," Proc. of Eurospeech-93, Vol.2, pp.1291-1294.

[6] Hosaka, J., Takezawa, T.(1992): "Construction of corpus-based syntactic rules for accurate speech recognition," Proc. of COLING-92, pp.806-812.

[7] Ehara, T., Ogura, K., Morimoto, T. (1990): "ATR Dialogue Database," Proc. of ICSLP-90, pp. 1093-1096.

[8] Fodor, J., Bever, T.(1965): "The psychological reality of linguistic segments," Journal of Verbal Learning and Behavior, pp. 4:414-420.

[9] Sugito, M.(1988):"Pause and intonation in discourse," Nihongo to nihongo kyouiku, Vol.2, pp.343-363 (in Japanese).

[10] Loken-Kim, K., Yato, F., et al.(1993): EMMI-ATR environment for multi-modal interaction. TT-IT-0081, ATR.

[11] Hosaka, J.(1993): A Grammar for Japanese Generation in the TUG Framework, Technical Report TR-I-0346, ATR.

[12] Sadanobu, T., Takubo, Y.(1993): "The Discourse Management Function of Fillers -a case of "eeto" and "ano(o)"-, Proc. of ISSD-93, pp.271-274.

[13] Hosaka, J., Takezawa, T., Uratani, N.(1992): "Analyzing Postposition Drops in Spoken Japanese," Proc. of ICSLP-92, Vol.2, pp.1251-1254.

[14] Kita, K., Kawabata, T., Saito, H. (1989): "HMM Continuous Speech Recogniton Using Predictive LR Parsing," Proc. of ICASSP-89, pp.703-706.