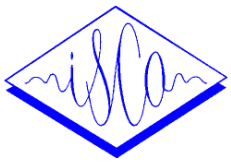

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A SPEECH AND LANGUAGE DATABASE FOR SPEECH TRANSLATION RESEARCH

Tsuyoshi Morimoto, Noriyoshi Uratani, Toshiyuki Takezawa,
Osamu Furuse, Yasuhiro Sobashima, Hitoshi Iida, Atsushi Nakamura,
Yoshinori Sagisaka, Norio Higuchi, Yasuhiro Yamazaki

ATR Interpreting Telecommunications Research Laboratories
2-2, Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-02, Japan

ABSTRACT

This paper describes a new database that we are constructing for speech translation research. A goal of the research is to develop a system capable of handling spontaneous utterances. The database is composed of three parts. The main one is an integrated speech and language database, which this paper mainly focuses. The basic approach for the database and its specific features are described. The other two, a speech database and a language database, supplement the main one. Brief descriptions about them also appear.

1. INTRODUCTION

Speech and/or language databases are important as basic material for speech translation research. Many kinds of speech or language databases already exist. We also have constructed such databases; a speech database [1] and a language database [2]. However, they were collected independently and their contents have little relationship with each other. The speech translation system we are planning to develop is one that can handle *spontaneous utterances*. Spontaneous utterances contain many new acoustic and linguistic phenomena which have been little understood.

This paper describes a new database that we are constructing for speech translation research. The database is composed of three parts. The main one is an *integrated speech and language database*, which this paper mainly focuses. The database contains not only speech wave data but also linguistic information such as morphological and syntactic tags. It also constitutes a bilingual corpus. All conversations take place between Japanese persons and English persons through human interpreters. We are planning to gather more than 1,000 simulated conversations, which will include more than 300,000 total words in both Japanese and English.

The other two databases, a speech database and a language database, are supplementary for this main one. Each database is constructed for a specific research purpose. The speech database contains monolingual conversations, and the language database does not contain speech wave data.

In section 2, our basic approach for designing a database for speech translation research is described. In section 3, the target tasks and domains of the database are mentioned. In section 4, the method used to construct the main database is introduced. Brief descriptions on both the supplementary speech database and the language database also appear.

2. BASIC APPROACH

Acoustic, linguistic and their related phenomena that would appear in spontaneous utterances, and that should be covered by the database, are as follows.

(1) Spectral and speaking rate variations

Spectral and speaking rate variations depend on the individual and speaking style and they are much broader than those of read speech. Unclear speech would be spoken, and its speed would change from time to time according to the intention of the utterance.

(2) Colloquial or situation-dependent expressions

Some colloquial expressions are hard to translate properly unless the dialogue situation of the moment is considered. For conversations concerned with complicated matters, a wide variety of speech acts, communication strategies and so on are used in the dialogue.

(3) Dialogue between people speaking two different languages

A language more or less reflects its culture. If the difference between two languages is large, such as between Japanese and English, it becomes very

difficult to translate the exact nuance of the input to the target language. The differences in communication strategies are similarly very large. For example, an ordinary Japanese prefers to use an indirect expression when s/he wants to request something, but a foreign (i.e., Western) person on the receiving end would become irritated, and vice versa. A professional interpreter can understand the underlying intention of the utterance, take account of the situation and provide a good translation of the expression in the target language. From this point of view, collecting bilingual dialogues in which two speakers with different languages and an interpreter between them is quite useful to know how the interpreter fills such gaps.

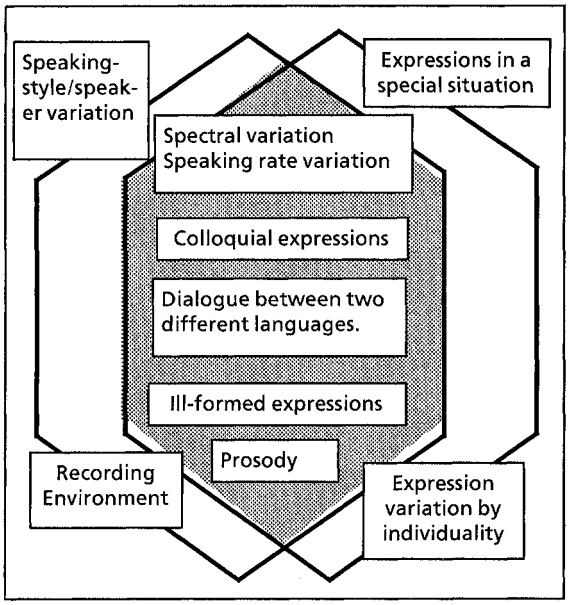

Fig. 1 Database for Speech Translation Research

(2) Ill-formed expressions

While spontaneously speaking, people think about how to say things or even what to say next. This sometimes produces several kinds of ill-formed expressions. Filled pauses and false starts are typical examples of such expressions. Investigating their characteristics is a quite important issue.

(4) Prosody

The prosody of speech carries a lot of intentional information of a speaker to a hearer. We believe that understanding a speaker's intention from both linguistic expressions and prosodic information, or controlling the prosody of output speech so that it reflects the speaker's intention is an important and attractive research theme for spontaneous speech translation.

All of the phenomena described above appear more or less as both acoustic evidence and language evidence. This is why we collect both speech and language in parallel and construct an integrated speech and language database.

3. TASK AND DOMAIN

We use a "task" and a "domain" having different meanings. "Task" means a purpose of a dialogue. For instance, both "inquiry for a hotel" and "inquiry for a flight" are the same task, i.e., *"inquiry"*. They differ in terms of the domain, i.e., application field. Thus, important parameter is "task". It is a measure of how complicated the conversation is. Our research goal is to develop a system in near future. As such, we do not want the target task to be too easy nor too difficult. From this point of view, we have chosen several tasks of different levels. The left-most column of Table 1 describes the tasks. They are listed in the order of difficulty.

As for "domain", we prefer a common one, that is, what ordinary people seem to have experienced and can easily imagine in terms of the conversational contents. From this point of view, we have selected "travel arrangement" domain. In this case, for instance, a conversation between a person and a travel agent or a hotel staff can be contained. In each entry of Table 1, examples of topics are described.

Table 1 Task and Domain

Task \ Domain	Travel Arrangement
• Simple Q/A	(not included)
• Inquiries & Reservations	• Hotel reservation • Inquiry and reservation for a tour •
• Scheduling • Suggestions	• Tour scheduling • Suggestion of shopping place •
• Claiming	• Troubles in payment • Claiming poor service •
• Complicated Conv. (i.e., trading negotiation or discussion)	(not included)

4. DATABASE CONSTRUCTION

4.1 Data Collection

Bilinguality is one of the most important features of our corpus. The main purpose of our corpus is to compile useful material for research aimed at implementing an executable system. Therefore, "sentence by sentence interpretation" is more reasonable than "simultaneous interpretation", which takes place through a simulated conversation. At the same time, this method of interpretation naturally limits the utterance time length to within a reasonable range (maximally, about 10 seconds).

Controlling the speaking turn is important in the same sense. If there is no control of speaking turn, a person tends to interrupt or throw in words of agreement frequently, and such interference sometimes causes incomplete sentences of the other speaker or overlapping of two speaker's utterances.

How to maintain the naturalness and reality of conversations and avoid inconsistent contents from appearing are other important issues while collecting simulated dialogues. One method we adopt is to prepare "a plot" for a conversation; this "plot" designates the detailed items to be asked or requested during the conversation. We also employ "a professional" on that domain. For the hotel conversation, for example, we nominate a person having much experience as a hotel clerk as a roll player. The number of subjects of our corpus is 50 for Japanese and nearly the same for English; they are all the native speakers of each language.

Preparing the plot, especially specifying the proper nouns such as human names or place names in the plot, is effective for preventing excessive divergence of the vocabulary size.

All conversations are done in a quiet studio environment, and recorded on a DAT. The data is digitized in 16 bits with 48 kHz sampling.

In the appendix, an example of a conversation is shown. The interjections are square bracketed. A capital letter J or E at the beginning of a line denotes a Japanese speaker or an English speaker respectively. The sentences in parentheses are the translations provided by the interpreter. The blank lines between utterances indicate turns of speaking.

4.2 Database Configuration

The collected conversations both in Japanese and English are transcribed as described above and these texts are stored in the database. The recorded speech wave data is segmented into utterances, linked from the corresponding text utterances, and then stored in the database. The correspondence relationships between each Japanese utterance and its interpreted English utterance as well as an English utterance and its interpreted Japanese utterance are indicated in the database.

Morphological analysis and syntactic analysis are performed for both Japanese and English texts semi-automatically and the resulting tagged data is also stored in the database (syntactic tagging of English texts is now under planning). Along with these data, linguistic resources used for the tagging, such as a word dictionary, morphological rules, syntactic rules and common tools are centralized and maintained uniformly. The configuration of the database is shown in Fig. 2.

5. SUPPLEMENTARY DATABASES

5.1 Speech Database

For speech recognition research, we are planning to gather a wider range of spontaneous speech. From a preliminary analysis of the above mentioned database collected so far, spontaneity of speech is affected by the constraints put onto it in some degree. In fact, we found that false starts or filled pauses, which are believed to be typical characteristics in spontaneous speech, appeared not so often, while the variation in the speaking rate was quite large as expected. One of the possible reasons for such unexpected results will be time delay between speakers alternating turns in the dialogue; this allowed the speakers to polish their utterances while they waited to respond. From this consideration, we are planning to gather monolingual conversations, in which there is no time delay due to interpretation.

We are also planning to extend the number and variation of speakers. At the same time, we want to cover various recording environments. Other conditions in collecting these speech data will be maintained; tasks, domains and controll of speaking turn.

5.2 Language Database

From our empirical studies on example-based machine translation [3] or stochastic-based parsing using a tree bank, we estimate that at least a one-million word corpus is necessary to obtain most linguistic expression pairs between a source language and a target language, as well as meaningful occurrence frequency of each word or some linguistic expressions, even when the target of the research is not spontaneous utterances. In spontaneous utterance research, a great variety of

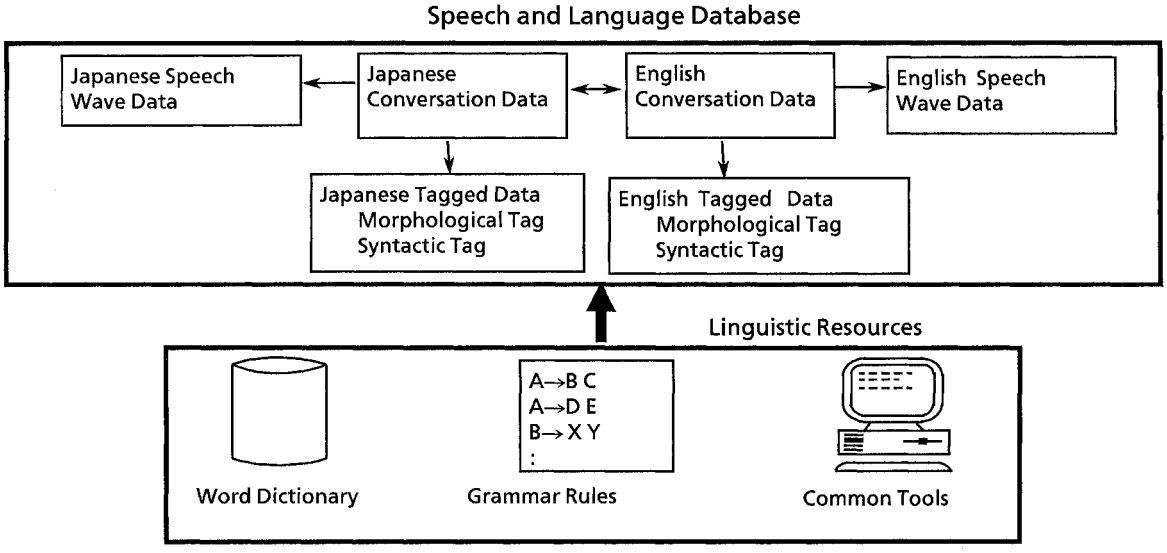

Fig. 2 Database Construction

communicative acts, utterance types, turn-taking patterns, communication strategies or global pragmatics will appear. Therefore, besides the integrated speech and language database, a bilingual language database with more than seven hundred thousand words on the same task and domain is being planned.

6. CONCLUSION

Our speech and language database for speech translation research has been described. It is composed of three parts; a integrated speech and language database, a speech database and a language database. Specific features of the first one are summarized as follows:

(1) It integrates speech wave data, transcribed texts and linguistically tagged data.

(2) It is a bilingual corpus. All conversations are made between a Japanese person and an English person through a human interpreter.

References

[1] Sagisaka, Y., Takeda, K., Abe, M., Katagiri, S., Umeda, T. and Kuwabara, H.: "A Large-Scale Japanese Speech Database," Proc. of ICSLP '90, pp.1089-1092, (1990)

[2] Ehara, T., Ogura, K. and Morimoto, T.: "ATR Dialogue Database," Proc. of ICSLP '90, pp.1093-1096, (1990)

[3] Sumita, E. and Iida, H.: "Experiments and Prospects of Example-Based Machine Translation," Proc. of 29th ACL, pp.185-192, (1991)

Appendix: A conversation example

E: Hello. Front desk.
(はい、フロントでございます。)

J; 五〇九<gomarukyuu>号室ですが、今テレビをつけてみたんですけど、全然映らないんですね。
(Hi, this is room five o nine. I turned on the TV right now but the picture doesn't show.)

E: I see.
(そうですか。)
E: Is the TV plugged in to an outlet?
(コンセントに電源は入っていますか。)

J: 電源は入ってます。
(Sure, it's plugged in.)
J: オンにすると一瞬つくんですが、パッと消えてしまうんですよね。
(I turned it on and it showed for a minute but it then suddenly disappeared.)

E: I see.
(分かりました。)
E: I'm afraid the repairman is gone for the day.
(あいにくですけれども、きょうはもう修理の者が帰ってるんですけれども。)
E: We won't be able to repair it until tomorrow.
(あしたにならないと修理できませんね。)

J: そうですか。それじゃあテレビは見られないんですか。
([oh] Does that mean I can't watch TV?)

•
•