THE VOICE ACROSS JAPAN DATABASE — THE JAPANESE LANGUAGE CONTRIBUTION TO POLYPHONE

Tom Staples, Joseph Picone, and N. Arai

Texas Instruments — Computer Science Laboratory
P.O. Box 655474, MS 238, Dallas, Texas 75265 USA
staples@csc.ti.com, picone@csc.ti.com, narai@trdc.ti.com

ABSTRACT

Texas Instruments' Voice Across Japan (VAJ) database, modeled after the highly successful Voice Across America project, consists of a wide range of diverse speech material including digit strings, yes/no questions, and phonetically-rich read sentences. The data is being collected using long distance telephone lines and an analog telephone interface. The target size is 14 items per speaker by 10,000 speakers. Greater emphasis is being placed on the collection of phonetically-rich read sentence data. Four randomly selected sentences are included in each session: one from the 512 sentence ATR PB set, and three from a 10,000 sentence set developed specifically for this project. This latter sentence set, designed to maximize the triphone coverage of the database, is described in this paper. VAJ is planned to be included in the Linguistic Data Consortium's (LDC) POLYPHONE (multi-language) database.

1. INTRODUCTION

The **Voice Across America** (VAA) database [1,2] represented the culmination of a two year effort to define and implement a comprehensive telephone database collection strategy. The VAA database was unique in that it combined extensive amounts of diverse speech material, ranging from digits to phonetically-rich TIMIT [3,4] sentences, detailed demographic profiles of each speaker, and qualitative assessments of the acoustic information into a single database. The result was a database that, even today, still supports fundamental explorations into the acoustic phonetics of spoken English in a real, operational environment — telecommunications.

The resulting VAA database consists of over 3700 speakers and over 50,000 utterances, and richly samples a host of demographic variables [2]. To date, we have collected over 10,000 speakers covering several domains using such database collection techniques. This type of database has proven invaluable for developing speaker independent speech recognition technology — especially technology based on phonetic modeling [5].

We are now in the process of replicating this data collection effort at Texas Instruments Tsukuba Research and Development laboratory. This program, called **Voice Across Japan** (VAJ), is being executed in collaboration with the Acoustical Society of Japan Database Committee [6] (a committee that is responsible for coordinating database activities in Japan). The goal is to collect a 10,000 speaker database.

The VAJ project was originally initiated as an internal project at Texas Instruments in 1989. Subsequently, the Linguistic Data Consortium (LDC) was created, and has established a strong initiative in multi-language databases. LDC plans to collect VAA-style databases on the order of 5,000 subjects and 200,000 utterances in several major languages (including American English, Spanish American, Japanese, Chinese, Dutch, and French). This project is known as POLYPHONE [7]. VAJ is intended to be the Japanese language contribution to this multi-language database. A core portion of each contribution to the POLYPHONE database will be similar. Hence, the database will support comprehensive cross-language experiments. In addition, spontaneous speech and some command/control words will be collected in certain languages.

Since the VAJ project was initiated prior to the creation of the POLYPHONE project, its design differs from the POLYPHONE specification in two significant ways. First, an analog telephone interface (similar to that used in VAA) has been employed. The POLYPHONE specification mandates the use of digital telephone interfaces (T1 or ISDN). Second, the number of utterances per session in VAJ is smaller. POLYPHONE recommends a session size on the order of 30 to 40 utterances. The reasons for these deviations will be discussed in the next section.

The POLYPHONE databases, unlike VAA, will be publicly available from LDC as per LDC's standard membership agreement. Hence, for the first time, significant amounts of data covering a common domain and a multiplicity of languages will be available. POLYPHONE will no doubt become the standard point of reference for multi-language and cross-language speech recognition.

0-7803-1775-0/94 $3.00

2. AUTOMATED DATA COLLECTION

The VAA database differed from all previous databases collected at Texas Instruments in that the data collection process was fully automated, with no human supervision of individual data collection sessions. Subjects called the system at their convenience and spoke a mixture of read and spontaneously elicited material. Previous databases collected over the telephone had required a trained supervisor to initiate and monitor each session. This is a costly and time-consuming process, typically requiring one month of collection per 100 speakers. Such a rate is clearly unacceptable in view of our goal of 10,000 speakers. Furthermore, this procedure is a poor simulation of potential applications, i.e. automatic telephone transactions, where machine-only interactions are the norm. These considerations led us to design and develop a robust system which could automatically handle the entire transaction.

The data collection front-end used in VAJ follows the model used in VAA. The incoming speech data is collected from an analog telephone line using a high quality telephone interface. The 2-wire to 4-wire conversion is performed using a Gentner TC-100 analog telephone line interface. Extremely high quality A/D and D/A is performed by a Sony ES series DAT, which is connected to the mid-impedance line-level audio inputs and outputs of the TC-100. The DAT data stream, which contains speech sampled at 48 kHz, is interfaced to the computer via a SCSI-based Townshend Computer Tools DAT-Link. The call processing portion of the TC-100 is interfaced to the computer using an RS-232 serial port. A Unix workstation (currently a Sun Sparcstation) controls the entire process.

The speech data itself is recorded at 8 kHz. The DAT-Link performs a high quality DSP-based downsampling of the data from 48 kHz to 8 kHz in real-time. Speech is automatically delimited from background noise during each transaction using energy-based endpointing. An utterance is extracted that includes approximately one second of channel noise preceding and following the speech data. Based on our experiences with VAA, in which only 0.5 seconds of background noise was retained, and based on the increased disk capacity available today, we decided to keep additional noise, so that the database would support robustness and noise modeling experiments.

Perhaps the most crucial decision in planning the database was the method of selecting participants. In VAA, a market research firm, the Home Testing Institute (HTI), was used to provide a demographically-balanced pool of participants. This subject pool can be classified as highly motivated for such volunteer participation tasks. Hence, response rate was very high — approximately 40%.

Unfortunately, no such panel exists in Japan. We were unable to identify a market research company that could provide us with such a resource. The next best alternative was to solicit employees of Texas Instruments in Japan (TIJ). Unfortunately, the TIJ employee population is heavily biased towards young male engineers. Hence, a strategy for better demographic coverage was required.

Each employee of TIJ receives a packet containing one cover letter explaining the project, and five session sheets, such as the one shown in Figure 1. Each TIJ employee is requested to recruit at least four participants — typically a spouse and two parents. Small gifts are included in the packet in an effort to encourage participation. TIJ contains approximately 5,000 employees; hence a total of 25,000 session sheets will be distributed.

The TIJ population, in addition to being age and sex biased, is also geographically biased (and consequently dialect biased). Hence, an important challenge for VAJ will be to overcome such biases. Several Japanese companies with much larger geographic coverage have expressed interest in providing access to their employee databases. Based on the incremental outputs of the validation process, we will use such companies to refine the solicitation strategy and improve the demographic coverage.

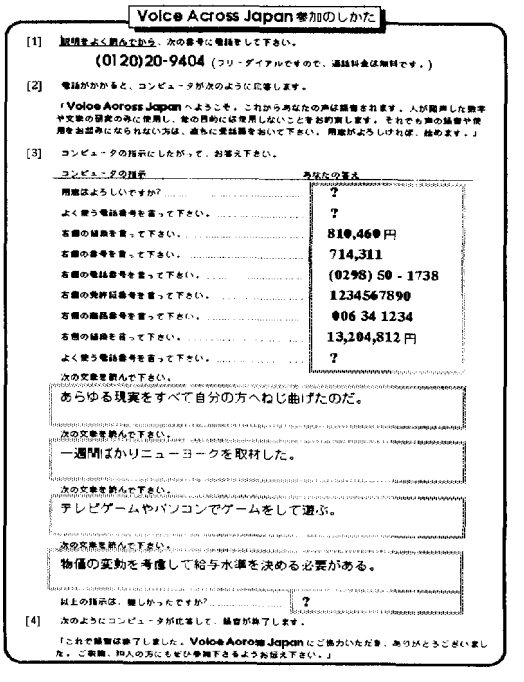

Figure 1. The Voice Across Japan session sheet contains fourteen items: two yes/no questions, two spontaneous spoken telephone numbers, six read digit strings, and four phonetically-rich sentences.

3. PHONETICALLY-RICH SENTENCES

One of the most significant changes since VAA in speech database research is the increased emphasis placed on phonetically-rich training material. The decision to add a TIMIT sentence in VAA was made as an afterthought — digits were clearly the focus of VAA. In VAJ, the phonetically-rich data is clearly the most sought after data. In fact, we are primarily interest in collecting data that gives a good triphone coverage and will support the development of context-dependent phonetic models.

As large as the VAA database was, it contained only 3700 repetitions of the TIMIT database sentences (we collected one sentence per session selected from an 1800 sentence subset of the TIMIT sentences — on the average two utterances per sentence for the database). Considering all aspects of acoustic variability (dialect and phonetic context are two of the more important), this amount of data is hardly sufficient for performing comprehensive experiments on context sensitive phonetic modeling. Hence, in VAJ, we are collecting four phonetically balanced sentences from each subject (we feel collecting a larger number of sentences from each subject is impractical given the current incentive/reward structure).

One of these four sentences will consist of a sentence randomly selected from the 503 ATR PB sentence set [8], since these sentences are considered a *de facto* standard in Japan, and have been used in numerous data collections. However, the ATR sentence set (as are most sentence sets) is too small for the needs of VAJ. This sentence set was carefully balanced by hiragana character pairs — most often a consonant-vowel (CV) cluster. The CV structure (a biphone context) is very important in Japanese — it is a fundamental unit in both the written and spoken language. The ATR sentence set, however, does not give as good a coverage of triphone contexts, mainly because of its limited size. Hence, we initiated a project to build a sentence set.

In the past, building such a database has been difficult due to the limited access researchers have to electronic text. Recently, however, a powerful new media containing easily accessible electronic text has appeared: the Sony Data Discman and its associated Electronic Book format. The Data Discman is a variant of the standard CD player that allows text data for be retrieved from the CD-ROM. We have downloaded the text in bulk from these discs onto our computers for analysis (the text, it turns out, is generally stored in a standard Japanese language format).

We have captured text from a wide selection of these discs: three sizable dictionaries containing words, representative sentences, and meanings; several discs containing reprints of back issues of famous newspaper articles; several discs containing language aides (common words, famous sayings); and several discs containing general information (travel guides, legal documents, etc.). Our raw text database is approximately 1 Gbyte of data, resulting in a database of over one million sentences.

These sentences were carefully processed using a procedure that consisted of selecting N phonetically-rich sentences using an entropy-balancing algorithm, reviewing each manually for suitability, adding the acceptable sentences to the database, and iterating until the desired number of sentences had been selected. Typically, at each step of the iteration, 1000 sentences were automatically selected from a pool of about 10,000 sentences, and 50% of these were manually rejected.

The entropy balancing algorithm used attempts to simultaneously maximize the entropy of several distributions. The Japanese writing system uses a system of over 8,000 common kanji characters. Since kanji is not a good predictor of the underlying pronunciation of a sentence, and hence its phonetic content, kanji was converted to hiragana before processing [9]. The entropy balancing algorithm operated on 1-grams, 2-gram, and 3-gram distributions of hiragana characters — often the equivalent of a 6-gram of phonemes.

A summary of the results of this selection process is shown in Figure 2. Entropies and the log of the number of n-grams are compared. We see that the entropy-balancing software does, in fact, produce a distribution better than that of either the ATR sentence set or a randomly selected sentence set. Note that the entropy of the randomly selected sentences is relatively high due to the fact that these sentences were not reviewed for suitability (which tends to

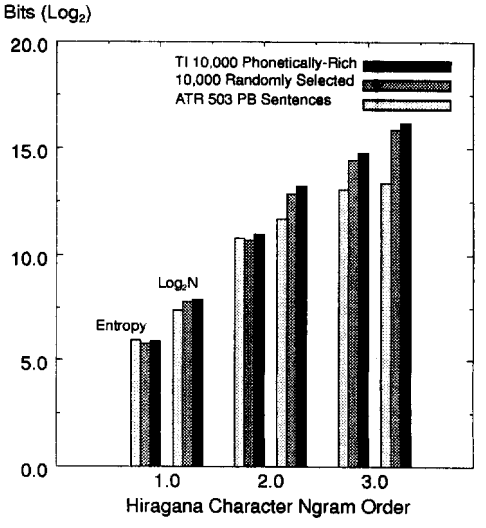

Figure 2. A comparison of entropy for several sentence sets is shown. Note that the TI phonetically-rich sentence set has the greatest entropy and the largest number of 3-gram contexts.

reduce the entropy of the set significantly).

Of course, part of the manual review process is to remove sentences difficult to read or to pronounce. In the case of Japanese, the writing system adds an element of difficulty. Some kanji characters are hard to read, occur infrequently, etc. Sentences were first reduced based on whether their characters fell within the Jōyō kanji character set [10] (a set of approximately 2000 common characters) and the number of characters in the sentence (sentences were constrained to be between 30 and 60 hiragana characters in length; 60 characters on the average is about 5 seconds of speech).

4. VALIDATION

Validation of VAJ will closely follow the procedure used in VAA [1]. Each file is orthographically transcribed, including transcriptions for non-speech events and other anomalous behavior. Several qualitative assessments of the data are made: signal condition, speaker effort level, articulation quality (normal/abnormal) and rate (fast/slow), speaker and signal quality (were there abnormal background noises such as echo?). Several qualitative assessments of the session are made as well (and shared by all files within the session): speaker sex, speaker age category, and most importantly, speaker accent. All this information, as well as the prompting text (the text contained in the session sheet), are stored in each utterance's data file.

5. STATUS AND PLANS

Initial review of the data and feedback from the pilot phases of data collection turned up several interesting differences from our experiences in the US. Perhaps the most worrisome of these was the fact that the average callers ambient environment is significantly more noisy than that in VAA. We hypothesize that this trend is a result of the smaller living areas and open work areas in found in Japan. For example, at TIJ, workspaces tend to consist of desks in large rooms, while in TI-US, higher-walled cubicles are more the norm. Similarly, the use of fashionable telephone equipment with open-air handset designs, and the central location of the telephone (near the kitchen and adjoining main living area) suggests a trend towards a more noisy ambient environment. Background television, radio, and multi-speaker noises appear in a much larger percentage of the pilot database sessions. We have taken additional steps in the instructions to encourage speakers to call from more quiet ambient environments.

As of the end of 1993, 500 speakers have been collected and are in the process of being validated. We expect data collection for the TIJ population to be completed by the first quarter of 1994. Validation of this phase of data collection will most likely be completed by the second quarter of 1994.

For further information on the plans and status of the distribution of POLYPHONE, contact Jack Godfrey, Executive Director, Linguistic Data Consortium, 441 Williams Hall, University of Pennsylvania, Philadelphia, Pennsylvania, USA 19104-6305, (Tel: 215-573-3595; email: jgodfrey@unagi.cis.upenn.edu).

6. REFERENCES

1. B. Wheatley and J. Picone, "Voice Across America: Toward Robust Speaker Independent Speech Recognition For Telecommunications Applications", *Digital Signal Processing: A Review Journal*, vol. 1, no. 2, pp. 45-64, April 1991.
2. J. Picone, "The Demographics of Speaker Independent Digit Recognition", in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 105-108, Albuquerque, New Mexico, April 1990.
3. L.F. Lamel, R.H. Kassel, and S. Senneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus, "in *Proceedings of the DARPA Speech Recognition Workshop*, pp. 100-109, Palo Alto, California, USA, February 1986.
4. W.M. Fisher, G.R. Doddington, and K.M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proceedings of the DARPA Speech Recognition Workshop*, pp. 93-99, Palo Alto, California, USA, February 1986,
5. Y.H. Kao, B.J. Wheatley, C.T. Hemphill, and P.K. Rajasekaran, "Toward Vocabulary Independent Telephone Speech Recognition," to be presented at the 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, Adelaide, South Australia, Australia, April 1994
6. S. Itahashi (Chairman), "Continuous Speech Corpus for Research," Committee on Speech Databases, Acoustical Society of Japan, vols. 1-3, ASJ-9101, January 1991.
7. J. Godfrey, "Preliminary Specification of the POLYPHONE Database," a personal communication to members of the POLYPHONE committee, May 7, 1993.
8. M. Abe, Y. Sagisaka, T. Umeda, and H. Kuwabara, "Speech Database User's Manual," ATR Technical Report No. TR-I-0166, ATR Interpreting Telephony Research Laboratories, September 1990.
9. J. Picone, T. Staples, K. Kondo, and N. Arai, "Kanji to Hiragana Conversion Based on a Length Constrained N-Gram Analysis," submitted to the *IEEE Transactions on Speech and Audio Processing*, Fall, 1993.
10. K. Lunde, *Understanding Japanese Information Processing*, O'Reilly and Associates, Sebastopol, California, USA, 1993.