CELLULAR-PHONE BASED SPEECH-TO-SPEECH TRANSLATION SYSTEM ATR-MATRIX

Rainer Gruhn, Harald Singer¹, Hajime Tsukada², Masaki Naito³, Atsushi Nishino⁴, Atsushi Nakamura⁵, Yoshinori Sagisaka, Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
E-mail: rgruhn@slt.atr.co.jp

¹ currently at SpeechWorks International, Boston, USA
² currently at NTT Service Integration Labs, Tokyo, Japan
³ currently at KDD R&D Laboratories, Saitama, Japan
⁴ currently at IME, Osaka, Japan
⁵ currently at NTT Communication Science Labs, Kyoto, Japan

ABSTRACT

We describe the implementation of a cellular-phone based speech translation system without telephone quality speech database or special CT hardware. The purpose is to quickly build a prototype service system that can be used for data collection with real users. To train the acoustic model for the speech recognition system, available high-quality databases were made usable by 1) appropriate downsampling and filtering of high-quality databases, and 2) by piping, similar to the NTIMIT and CHIMIT paradigms. An evaluation of acoustic models with filtered, piped and real cellular-phone data is given. Recognition rates are at same levels as for wideband speech.

1. Introduction

For more than a decade, ATR has been carrying out research in the field of speech-to-speech translation technology. Our most recent research prototype system, ATR-MATRIX, can translate Japanese speech into English, German, Korean and Mandarin, and English speech into Japanese [1].

Two key issues should especially be considered when aiming at a system which can be used in the real world. One is to make the system more robust against the variability in human speech, environmental non-speech sounds and so on, that are typical problems in the case of using a spoken language system in a real environment. The other issue is to make the system easily accessible from anywhere. Since conventional speech-to-speech translation systems, including ATR-MATRIX, have been available only on expensive and heavy high-end computers, only a limited number of people can use the system at limited places. One solution to this is to down-size the whole system by “pruning” the speech translation software, and by further progress in the integration technology of processor and memory devices. The other solution is to split the system into two parts, a computationally expensive part running on server computers in the network and a (minimal) speech input function part running on the user’s laptop or hand-held terminal. This architectural solution allows the user’s cost to be minimized without sacrificing performance compared to the standard ATR-MATRIX.

Several implementations to split the system are possible. In [2], we proposed a design in which feature extraction for speech recognition and wave unit concatenation for speech synthesis are running on the user’s laptop. In this paper, we discuss a design in which the user only needs a cellular-phone to access the speech translation system. In Japan, cellular-phones are already used by more than forty million people and they have a very vast access area. A system based on the cellular network enables the speech-to-speech translation system to be easily accessible anywhere. The translation server can be located anywhere and is connected to the telephone network using some CT (computer telephony) interfaces.

This paper describes the setup of our cellular-phone based prototype, a procedure for training/adaptation of acoustic models for cellular phone speech by re-using previously collected databases, and remaining major problems in this design towards a real-world speech translation service. Future work will include using the prototype system to collect data and use this data to improve it towards a real service system.

2. System Description

![ATR-MATRIX system structure](image)

Figure 1. ATR-MATRIX system structure

The speech-to-speech translation system ATR-MATRIX was designed with a modular structure, consisting of the speech recognition subsystem ATRSREC [3], the language translation subsystem TDMT [4], the speech
Table 1. Accuracy (%) and rtf (real-time factor for a Pentium II-300MHz) for acoustic models adapted with different data.

<table>
<thead>
<tr>
<th>evaluation speech</th>
<th>model type</th>
<th>word accuracy</th>
<th>phoneme accuracy</th>
<th>rtf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wideband</td>
<td>Wideband</td>
<td>89.0</td>
<td>79.3</td>
<td>0.45</td>
</tr>
<tr>
<td>Keitai PHS</td>
<td>Baseline</td>
<td>87.3</td>
<td>69.1</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>trueKeitaiPHS</td>
<td>88.8</td>
<td>71.1</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>1cmKeitaiPHS</td>
<td>89.3</td>
<td>69.7</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>2cmKeitaiPHS</td>
<td>84.1</td>
<td>66.0</td>
<td>1.02</td>
</tr>
<tr>
<td>Keitai Keitai</td>
<td>Baseline</td>
<td>83.1</td>
<td>68.8</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>1cmKeitaiKeitai</td>
<td>89.8</td>
<td>74.8</td>
<td>1.03</td>
</tr>
</tbody>
</table>

synthesis subsystem CHATR [5] and a main controller (see Figure 1).

Figure 2 gives an overview of a bi-directional ATR-MATRIX with cellular-phones. It consists of two Compaq XP1000 workstations with Alpha 21264 500MHz processor and 368MB memory, running Digital UNIX V4.0.E. Four cellular-phones (DoCoMo Doccimo SH311) are used, two on the server side and two for the customers. The headset plug of the cellular-phones is the interface to the computer. The cellular-phone output (i.e., the speech from the client phone) is connected to the line-in port of a DAT walkman. The input plug is connected to the soundcard’s speaker port on the computer. Figure 3 shows a closeup on the ports and connections. The speech recorded on the DAT is both fed into the speech recognizer and saved on harddisk.

Most telephone based speech recognition systems use special CT boards on the server side. For a commercial service, these CT boards are indispensable. However, in this work we choose not to use a CT board because of:

- development time and cost,
- availability only for certain platforms,
- usually only half-duplex,
- input speech is plan-encoded.

In Japan, two cellular-phone types are common, “Keitai”\(^1\) using PSL-CELP speech encoding and “PHS”\(^2\)

\(^1\)“Keitai” means “portable” in Japanese 
\(^2\)“PHS” stands for Personal Handyphone System


using ADPCM (24 kbps), Keitai and PHS are independent network systems. The cellular-phones we use instead of CT boards can support both, making experiments with the different channel properties easier. They have two telephone numbers, one for Keitai and one for PHS. When making a call, the transmission system can be chosen freely, when receiving, the system is chosen based on which of the two telephone numbers was called.

For control access like onhook, offhook or dialing we used the modem port of the cellular-phones, which is possible even during a normal (voice) call. As the modem card drivers were only available for Microsoft Windows but the ATR-MATRIX server machine was a Unix system, a separate PC running Windows was used as modem control device in our prototype translation service system.

During our experiments we put the cellular-phone in the battery loading station to ensure continuous power supply. There was no noticeable impact (like network hum) on the audio quality.

The cellular-phones used for our experiments are slightly modified to additionally connect to ATR’s office-internal cellular-phone network which is also PHS based. The ATR building appears to be far from the next cellular-phone network node so that especially in windowless rooms in the center of the building conditions for normal cellular-phone connections are unreliable.

We concentrated our efforts on connections where the

figure 3. Using a cellular phone instead of a CT board in a cellular-phone based service system.
customers use a Keitai. The cellular-phones on the server side work in office-internal, PHS-based mode (keitaiPHS). An ATR-MATRIX system with Keitai both for server and for client has also been implemented (KeitaiKeitai). The assumption that most clients have a Keitai was made because Keitai type cellular-phones are more widespread than PHS cellular-phones and because the area coverage of the PHS network is smaller.

The connection between the two ATR-MATRIX servers follows the THOP protocol (AT&T proprietary) which is based on SIP [6]. The transmitted data is translated text. The receiving system generates synthesized speech.

3. Data Collection and Evaluation

For any new acoustic condition and task domain, matched data has to be collected to train or adapt the statistical acoustic and language models. At ATR, we have been collecting huge databases for high-quality speech [7, 8]. However, we had no access to cellular-phone and not even telephone speech databases in similar task domains.

We thus experimented with several methodologies to reuse ATR’s wide-bandwidth databases:

1. downsample to 8kHz and apply telephone FIR filter (gained from an ITU data sheet). The band response curve is shown in Figure 4

2. pipe speech data through a speaker – cellular-phone connection and re-record, similar to NTIMIT and CTIMIT[9] as shown in Figure 5. Unlike CTIMIT, we piped the database making only a single call and in an (except for air condition) quiet office room.

![Figure 4. CCITT Telephone Band Response Curve](image)

3.1. Database Description

**set1**: 166 male and 241 female speakers, each spontaneously uttering one conversation side in the travel arrangement task, a total of 6432 utterances, recorded at 16 kHz with a Sennheiser headset.

**set2**: 10 male speakers reading 50 phone-balanced sentences each, collected simultaneously with 2 channels, one at 16 kHz with a Sennheiser headset and one at 8 kHz through a keitaiPHS connection.

**set3**: 5 male speakers reading 48 travel arrangement sentences each, collected similar to set2.

We also recorded **set4**, similar to set3, except that we used KeitaiKeitai connection. The high-quality recording of **set2** was then piped through a keitaiPHS (keitaiKeitai) connection, varying positions between playback speaker and cellular-phone microphone.

3.2. Training, Adaptation and Evaluation

**Set1** was downsampled to 8 kHz and the FIR filter applied. The resulting data was used to train a gender-independent state-sharing HMM topology with ML-SSS[10] for a 1000 state model. Gender-dependent models were then retrained using Baum-Welch algorithm and the number of mixtures at each state increased to 5. We refer to this acoustic model as “Baseline AM”.

We then performed MAP-VFS environment adaptation [11] on the Baseline AM using channel 2 of **set2** and denote this as “truekeitaiPHS AM”. Alternatively, piped data of **set2** was also used with MAP-VFS at both 1cm and 20 cm distance between speaker and cellular-phone microphone (1cmkeitaiPHS and 20cmkeitaiPHS AM, respectively).

Continuous word recognition was performed with a 2-pass decoder with a vocabulary size of about 13,500 words using a multi-class composite n-gram[12]. For comparison, we also performed phoneme recognition using a phoneme bigram trained on the 407 training conversations.

3.3. Discussion of Results

Table 1 shows the evaluation results. The most surprising result was, that wideband speech does not outperform the cellular speech (89% vs. 89.3%). For cellular-phone speech, we had a considerable drop in phoneme accuracy, from 79.3% to about 70%, but the word accuracy remained
at about the same level or even rose slightly. From this we can see the strong influence of the language model on recognition accuracy. The strongest impact the lower data quality has is on the real-time factor, which doubles compared to wideband speech.

The difference of KeitaPHS and KeitaKeita speech becomes visible in the performance of the Baseline acoustic model, which performed rather poor on KeitaKeita speech (83.1%). Piping speech proved to be a cheap and fast way to generate data for an acoustic model that works very well in a new scenario. An important issue in piping is to set up the experiment properly, e.g. distance and position of speaker and cellular-phone have an impact on performance.

A different approach that we tried was using time-stretched pulse signals (TSP) [13] to measure acoustic channel characteristics. We piped several TSPs and then filtered the wide-bandwidth database with the averaged impulse response. Accuracy for acoustic models trained with that method was less than 50%, i.e. a huge mismatch.

4. Remaining Problems

- **Environmental Noise**: Microphones in cellular-phones pick up a lot of environmental noise. To improve the input device of cellular-phones is indispensable for a speech recognition service application. Environmental noise causes start-and-end point detection (EPD) errors on the input speech as well as acoustic model mismatch. It is necessary to develop noise-robust methods for both EPD and acoustic modeling.

- **Dropout**: In current cellular-phone services, sometimes short transmission interruptions occur. Even for a human it is difficult to understand utterances with dropouts. More reliable cellular-phone services are desirable for speech recognition applications.

- **Design of Human Interface**: In our prototype implementation, users are not informed a) if the system is currently processing, b) if the recognition subsystem rejects the input speech, and c) if the recognition/translation results are correct. As for (a) and (b), the system can give such information to users by playing music etc., whereas c) is a fundamental problem for a speech-to-speech translation system. Our current design strategy of the system is to do nothing about (c) since users probably get feedback through the conversation. However, it would be worth developing methods to know (c), for example using paraphrasing or retranslating into the source language.

- **Name Recognition**: As the target topic for our system is hotel reservation and travel information, person and place names play a very important role [14]. As the variety of names is too high, reliable unconstraint name recognition will not be feasible in the near future. One possible approach to avoid name recognition (and translation) is to directly transmit the untranslated source speech to the dialog partner.

5. Conclusion

We have implemented a cellular-phone based speech translation system and shown several approaches to reuse previously collected high-quality data. For a real-world service, however, human-interface and noise robustness issues still require a major research effort.

References


