Progress and Prospects for Speech Recognition in Broadcasting

1. Introduction

Speech recognition technology, which automatically generates a textual script for human speech, or allows a user to operate a device with verbal commands, has significantly benefited from recent advances in statistics, computing, and large-scale speech and text databases. The Science & Technical Research Laboratories (STRL) initiated research on phoneme recognition of the five Japanese vowels over thirty years ago, followed by research on number recognition, speaker-dependent phrase speech recognition, and speaker-dependent large-vocabulary continuous speech recognition. NHK is currently providing closed-captioned broadcasting using speech recognition. The service started in March 2003, on the NHK program "News at 7" for the segments where an announcer read from a manuscript in a studio. This service was later expanded to "News at 9," "News at Noon," and "Good Morning, Japan." The subtitling of other live broadcast programs employs a re-speak method in which the recognition system recognizes, not the program audio directly, but a separate speaker called a closed-captioning announcer, who listens in a quiet room to the original audio and repeats aloud for processing. This method has been used since the end of 2001, in live broadcasts of "Kohaku Utagassen" (a New Years Eve music show), the Olympics Games, the Grand Sumo tournaments, and professional baseball games.

The common method of using a special keyboard for high-speed transcription to provide real-time captioning in Europe and the United States has not worked well for the Japanese language, which requires conversion from Kana to Kanji characters. Thus, speech recognition technology in Japan can be considered the driving force for closed-captioned live programs, which have long been anticipated by those with hearing impairments. NHK has steadily expanded its closed-captioned service, with the percentage of closed-captioned programming for General TV broadcasting in fiscal 2003 reaching 33.8%.

Besides closed-caption broadcasting, there are other promising broadcasting applications of speech recognition. Advanced broadcasting based on home servers, which is a future integrated services television system, will require a diverse range of program-related information called metadata to enable the user to retrieve specific program scenes or desired pieces of information. Speech recognition technology is expected to facilitate easy extraction of speech content from a program's audio data based either on the casts' utterances or keywords, by detecting information relevant to a scene's content, such as who said what at what time. At present, studies aimed at efficient metadata generation are focused on speech recognition, and we have targeted Major League baseball in the United States and soccer programs. Advances are also being made on a spoken-dialogue directive television interface that is easy for anyone to use.

While speech recognition technology can now be used to realize an efficient man-machine interface, it still faces significant problems for closed-captioned broadcasting or for generating metadata from various program sources. For instance, a word recognition accuracy (the ratio of accurately recognized words out of the total number of words spoken) of approximately 98% can be achieved when a speaker with clear pronunciation reads from a manuscript in an environment with no background noise. However, the following factors adversely affect recognition accuracy in less controlled environments.

- Background noise such as traffic noise and cheering.
- Unclear pronunciation.
- A colloquial style of speaking, or casual content, like in everyday conversation.
- Content that is outside of the assumed recognition target.

Because of these factors, present speech recognition technology has not yet achieved a level to be easily used in normal acoustic environments or with a variety of program content.

In the next section, we discuss speech recognition technology for broadcasting, its research agenda, and prospects of reducing the influence of background noise.
2. Transcribing Broadcast News

2.1 System overview

A broadcast news transcription system based on speech recognition automatically recognizes an announcer’s voice while he or she is reading from a manuscript in a studio, and generates a script for closed-captioning after prompt manual correction of any recognition errors by the system (Figure 1). A probabilistic statistical method of speech recognition is generally employed to check the word appearance frequency in reporters' manuscripts learned in advance, or the characteristic distribution of an announcer's voice, to determine the word sequence with the closest match to the input speech. The system performs: 1) sequential extraction of acoustic features, such as frequency characteristics and 2) matching of the input speech acoustic features with the acoustic models, which reflect the pronunciation of all the registered words (approximately 20,000 words for news). An acoustic model statistically approximates acoustic feature distributions by phoneme based on the speech database; it does so to compute the acoustical similarity of the input speech to all the phonemes at the time of recognition. This acoustic score calculation is followed by, 3), reference to the language model (sequential word appearance frequency distribution) at the time of processing the transition from one word to the next to store only high probability (language score) words as candidates. The last stage, 4), uses the aforementioned procedure until utterance completion; the output word sequence has the highest probability, both acoustically and linguistically (the maximum acoustic and language score combined). In addition, to make closed-captioning after error correction as prompt as possible, word candidates with adequate reliability are output even before the speaker has finished speaking.

Recognizable expressions are automatically incorporated from reporters’ manuscripts right before each broadcast, to prepare the system for new proper nouns and topics that may appear. The system has also been given acoustic models of many NHK announcers, so it can recognize many unspecified speakers. Two separate teams, each comprising an error detector and an error corrector, perform real-time error correction of the results. The average delay between when the announcer completes a speech segment utterance and when the caption appears on the TV screen is approximately 10 seconds. This method is not adequate for field reports or conversations, and subtitling of such segments is done by operators using a special high-speed transcription keyboard. The effects of background noise and speech style thus remain significant research issues.

2.2 Present system performance

Table 1 shows the recognition accuracy if target news program speech is categorized by acoustic feature, such as by speaker and background noise, and by linguistic feature, such as by whether it is a manuscript being read or a conversational speech. Manuscripts read out in a studio are both acoustically and linguistically advantageous, and with them, we have achieved approximately 98% recognition accuracy. However, mismatching with the learned linguistic model increases when the target speech is field reports, dialogue and commentary, or completely spontaneous. The adverse

<table>
<thead>
<tr>
<th>Speech category</th>
<th>Acoustic feature</th>
<th>Linguistic feature</th>
<th>Recognition accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript readout in studio</td>
<td>Announcer: No noise</td>
<td>About the same as manuscript</td>
<td>Approx. 98%</td>
</tr>
<tr>
<td>Material from local broadcasting station, completed production material</td>
<td>Announcer, Reporter: Noise</td>
<td>About the same as manuscript</td>
<td>Approx. 96%</td>
</tr>
<tr>
<td>Field reporting</td>
<td>Announcer, Reporter: Noise</td>
<td>Close to manuscript, with additional interaction</td>
<td>Approx. 91%</td>
</tr>
<tr>
<td>Dialogue/Commentary</td>
<td>Announcer, Reporter: No noise, Rapid speech rate, Unclear pronunciation</td>
<td>Outline and memo used, but generally conversational spontaneous speech</td>
<td>Approx. 82%</td>
</tr>
<tr>
<td>Interview</td>
<td>Politician, General: Noise</td>
<td>Spontaneous speech</td>
<td>Approx. 45%</td>
</tr>
</tbody>
</table>

Table 1: News program speech types and current recognition accuracy
Acoustical features for such speech include unclear pronunciation and rapid speech when the speaker shifts from announcer to reporter, and then to another speaker. Other acoustic factors are background noise and differences in circuit characteristics during telephone reporting.

A news program announcer speaks approximately 3.3 words on average per second. Thus, a minimum recognition accuracy of 95% would involve real-time manual error correction of one word per six seconds on average. Past experiments have shown that 90% recognition accuracy makes real-time error correction impractical. Out of these, recognition during interviews including background noise and spontaneous speech by ordinary speakers is considered extremely difficult.

2.3 Research Agenda

2.3.1 Acoustic features

Speech recognition does not deal with the temporal waveform of the speech itself, but with the acoustic features derived from only the components relevant to the speech content extracted from the temporal waveform (acoustic analysis). The frequency characteristics of the input speech (lower DCT primary component in Figure 2 spectrum) are commonly used for this purpose, and a lot of researchers have sought acoustic features that give higher recognition accuracy. Especially sought after are acoustic features that are robust against fluctuations related to a noisy environment and speech style and those that can suppress noise components.

We have tried to extract independent temporal characteristics, rather than the conventional frequency characteristics, over multiple frequency bands, and to calculate more detailed dynamic acoustic features equivalent to the dynamic frequency characteristics. Our research led to higher accuracy for field reporting and dialogue segments. A promising noise control is the filterbank subtraction method, in which additional noise is suppressed in multiple frequency bands independently. Moreover, joint research between NHK and BBN Technologies in the United States confirmed that normalizing multiplication noise in the circuit or microphone improves recognition accuracy for news audio.

The acoustic features are extremely important in the initial stage of speech recognition, yet they are severely influenced by background noise, circuit distortion, and pronunciation disturbances. It is hoped that more robust acoustic features can be found by exploiting unconventional information, such as supplemental acoustic information dealing with pitch and speech rate, or auditory psychological characteristics information on masking and illusion.

2.3.2 Acoustic model

An acoustic model approximating the acoustic features distribution by phoneme (Figure 2) is usually built from speech similar to the recognition target, because the closer the training conditions are to actual recognition conditions, the higher the recognition accuracy will become. These conditions include the speaker's vocal characteristics, clarity, speech rate, speech style, and background sound level and type. Ideally, the acoustic model can be adapted using a smaller amount of speech data if the data are close to the recognition target.

It may also be possible for the recognition system to use multiple acoustic models adapted to acoustic conditions, rather than using multiple recognition systems, if the optimum adapted acoustic model can be automatically selected for the input speech.

As mentioned above, NHK together with BBN Technologies have developed a method of configuration by phoneme unit that consists of

![Figure 2: Acoustic features and acoustic model](image-url)
To improve the language model’s performance, we are using the word sequences that appear in the latest news manuscripts, as well as language model adaptation to speech style and a technique to enhance the statistical confidence measure by adding frequently used compound words to the vocabulary. We are also studying how to integrate multiple confidence measure scales, such as an acoustic score to learn recognition error tendencies.

Covering the broad range of topics in broadcast programs requires a several hundred thousand word vocabulary. Moreover, besides word sequence appearance frequencies, the recognition system will need knowledge of the meaning, content, and grammar. To this end, we will devise a system that incorporates thesaurus and word class information reflecting the word’s meaning.

2.3.4 Correct word search

A correct word search is a process in which an acoustic score is obtained by matching the acoustic features with an acoustic model, and adding a language score for sequential words, to determine the most probable word sequence. Since a meticulous score computation for every possible word sequence combination, i.e., a full search, is virtually impossible, our decoder searches for and stores possible word sequence combinations, i.e., a full search, is virtually impossible, our decoder searches for and stores possible word sequence combinations, i.e., a full search, is virtually impossible, our decoder searches for and stores possible word sequence combinations, i.e., a full search.

The broadcast news transcription system has four operators, two error detector and error corrector pairs. In view of securing and training staff, a smaller number of operators would be desirable. Thus, we did a study showing that the work of the four operators could be done by two persons if a new error correction system was adopted. We are also conducting research on automatically estimating recognition errors to support manual correction work. Real-time closed-captioning requires not only the speech recognition but also efficient error correction. For this we are studying feeding back the error correction results to the speech recognition system.

3. Re-speak Method Speech Recognition

3.1 System overview

The re-speak method is employed in live broadcast programs other than news programming for subtitling production. The re-speak strategy recognizes speech repeated by a separate speaker, called a closed-captioning announcer, rather than speech from the original program’s audio (Figure 3).

The closed-captioning announcer repeats aloud what announcers and performers say in the program, or summarizes its content if necessary, as he or she listens to it on headphones. Specific acoustic models are used for every closed-captioning announcer. The system switches to the acoustic model when the closed-captioning announcers change, and the language model contains data specific to the program’s genre, such as the type of sport or music talked about. The language model is tuned before each broadcast with additional textual information, such as the program script, anticipated proper nouns including personal and geographical names, and technical terms. Live sports broadcasts, in particular, require prompt display of closed-captions, so in such cases, one operator prioritizes the deletion of recognition errors over correction. Individuals with hearing impairments have expressed to us that this form of closed-captioning has increased the information and satisfaction that they get from music and
sports programs.

The re-speak method can reduce acoustic and linguistic problems in a program's audio, such as noise, overlapping voices, and casual conversational expressions. It enables a closed-captioning announcer to summarize or supplement program content, to make the closed-captioning easier to understand, and to add supplemental information such as clapping and cheering (Table 2).

3.2 Current performance

The re-speak method is now being used for various sports and music programs produced by NHK, achieving approximately 95% recognition accuracy. On the other hand, the recognition accuracy for the lifestyle program “Fresh Morning” remains at approximately 90%. The reason for the lower rate is lack of relevant learning data, as the program covers an extremely broad range of topics, such as cooking, health, and hobbies. Table 3 compares the performances of the language models used for “speed skating” at the Winter Olympics and for “Fresh Morning.” Perplexity is an index to describe the word estimation capability of a language model (equivalent to gauging how many possible words were considered to estimate one word), with a lower value indicating higher performance. The trigram hit rate is the frequency with which a three-word sequence registered in a language model appears in the evaluated program (higher value = higher performance). The out-of-vocabulary (OOV) rate shows how many words in the evaluated program are unregistered in the language model vocabulary (lower value = higher performance). These three indexes collectively reveal that language model learning is not at an adequate level; specifically, the word estimation for “Fresh Morning” is about twice as difficult as for “speed skating,” with approximately 10% fewer three-word sequences included in the language model, and 10 times more unregistered words in the vocabulary.

3.3 Research agenda

Building a system that can recognize a broad range of topics, as might be found in lifestyle shows, will be a challenge because even a small topic category may require a large learning text. An adequate learning method would be one in which basic language models are separately compiled by sub-genre and then efficiently used to adapt the language model of the program. The basic language model would have a super-large vocabulary made from a large vocabulary database for use in general broadcast programs and would not use the conventional vocabulary building design based on word appearance frequency in the learning text.

Although the current system uses the re-speak method, our ultimate goal is direct recognition of a program’s audio
content. By overcoming the issues related to background noise, unclear pronunciation, diverse expressions in a spontaneous conversational speaking style, and the overlapping voices of multiple speakers we hope to achieve a system capable of direct recognition for any type of broadcast program.

4. Metadata generation

Future broadcasting based on home servers will transmit program related-information called metadata along with regular programming, and efficient metadata generation is the key to enabling a wide range of new viewing styles. STRL has been studying many methods to generate metadata automatically. The 2004 STRL Open House included an exhibition on keyword extraction of baseball terminology and player names from announcer’s speech integrated with face recognition and scene analysis during a Major League Baseball game (Figure 4).

The acoustic and language models were adapted to specific announcers and commentators by analyzing past program audio data of Major League and Japanese professional baseball relay broadcasts, as well as their transcripts. Noise countermeasures were used to get the acoustic score in segments with a lot of crowd noise. These measures improved the speech recognition accuracy and extraction rate for keywords (baseball terms, player names, geographical names, etc.). Yet little enhancement was achieved for segments featuring conversations with a commentator (pointing to both a lack of pronunciation clarity and a deficiency of relative learning data as problems), showing the large influence of spontaneous speaking.

Model performance may be improved by identifying the speaker and background noise and by examining ways to exploit information such as error tendency and word meaning. Metadata related to scene content might also be used to improve speech recognition performance. We plan to study the word types that prove useful for metadata production, especially which words should be avoided to reduce recognition errors. We will also look into integration of other automation technologies including face recognition and language processing for closed captioning.

5. Conclusion

How will speech recognition technology evolve as a program production support tool for closed-captioning and metadata production in the future? While current research centers on problem solving through the use of features for specific applications, our goal for the next 10 years is the construction of a general-purpose super-large vocabulary continuous speech recognition system for any application. This will be a single speech recognition system for closed-captioning, generating metadata, editing support, and inputting system commands that can recognize target speech of various programs, materials, and operators. It is expected that this speech recognition system will evolve into a general-purpose, distributed speech recognition server that can return recognition results for speech entered over a network.

The realization of such a speech recognition system will require spoken language processing that is far more advanced than the current level. For example, present-day speech recognition technology is based on word
appearance frequency; it does not extensively use meaning, context, or grammar that humans use to understand speech. Yet a continuous speech recognition system with a several hundred thousand word vocabulary to cover the diverse topics and vocabulary of broadcast programs will be achieved only through the mastery of such extensive knowledge. Present audio processing is based only on a rough representation of the frequency characteristics of a voice, falling far short of making full use of the information that humans use in daily life, such as supplemental acoustic information dealing with pitch and speech rate, or the auditory psychological characteristics of masking and a knowledge of illusion. The speech recognition system should be improved by such advanced acoustic analysis based on the fact that humans can easily understand speech in noisy environments and colloquial speaking styles. We will raise our speech recognition technology to a higher level by the advanced acoustic and linguistic processing in order to utilize it in general applications including closed-captioning and metadata production.

(Toru IMAI, Associated Director, Human Science
Takayuki ITO, Director, Human Science)

1 A technology that covers a vocabulary of over 10,000 words, and recognizes continuous speech pronounced without word divisions.
2 The ratio of closed-captioned broadcast programs that can be subtitled, out of the total broadcast time, excluding programs such as those that are live, is 92.4%.
3 Manuscript readout also involves revision of the original electronic manuscript. Announcers frequently summarize the manuscript, change the expressions, or construct new sentences by combining multiple scripts in a program.
4 The Discrete Cosine Transform (DCT)
5 A 10-word-speech using a 20,000 word vocabulary database has $10^{43}$ possible combinations.