A Framework of Human-based Speech Transcription with a Speech Chunking Front-end

Takashi Saito
Shonan Institute of Technology, Kanagawa, Japan
E-mail: saito@sc.shonan-it.ac.jp Tel: +81-466-30-0188

Abstract— **This paper presents a framework of "human-based" speech transcription in a crowdsourcing environment. The main purpose of the framework is to promote participation of a large population of volunteers in speech transcription to create caption data for hearing-impaired people. It allows volunteer participants to join the transcription task with a very short segment of speech, called here as "speech chunk". It is realized by effectively incorporating a front-end of speech chunking prior to the main transcription task. The front-end is intended to increase the flexibility of the transcription task allocation to participants and more importantly to reduce the burden of the task itself by chopping audio data in advance into appropriate length of utterances and accordingly easing the repetitive playback operations. As an initial study, the performance of the speech chunking is investigated for various types of contents on how appropriately speech chunks are extracted as a transcription task unit. The result shows that the framework can be applied even to animation video contents that usually include dynamic sound effects.**

I. Introduction

The number of digital contents such as video, movie, or podcast has been continuously growing on the broadband internet. Even though the internet has become one of essential channels to obtain daily useful information, the rate of captioned digital contents is still low on the internet. Captioned data is indispensable information to people with hearing impaired, and is also beneficial to aged people who are getting hard of hearing, or foreigners who are not familiar with the language of their visiting country. The low rate is primarily caused by the fact that the captioning work, that is, speech transcription, is a labor-intensive and consequently expensive task. The use of automatic transcription by speech recognition is also limited because of its accuracy insufficient to real-world contents in various audio conditions.

Transcribing speech is intrinsically an easy task for humans who can hear and write in their native languages, and the performance quality can be kept good if only the work volume is reasonably small. Therefore, one of the promising solutions is to utilize the crowdsourcing mechanism to effectively collect contributions of a large population of people to the speech transcription problem. In the field of speech transcription research, there are number of previous studies reported to make use of human contributions. In [1], a collaborative editing facility for correcting speech recognition errors was developed in a Webcast viewer of university lecture videos. Assumed participants are mainly students in the course. PodCastle [2] involves much broader participants of anonymous users on the internet. The task is to correct errors of speech recognition applied to podcasts. In [3], groups of non-expert captionists (people who can hear and type, but are not trained stenographers) contribute to real-time captioning from scratch. The system automatically merges partial input captions from participants into a single output stream. More generally, as the form of speech crowdsourcing, the overall trend and future challenges are discussed in [4].

This paper focuses on how to reduce the task burden of participants in a crowdsource-based speech transcription framework, which is intended to promote participation of volunteers in speech transcription to create caption data for hearing-impaired people. In a common style of speech transcription task, participants need to repeat the basic cycle of transcription: setting the audio position, playing back, and typing in text. The work burden coming from the simple repetitive operations may not be neglected. The less the burden is posed, the more volunteers are expected to easily participate from dedicated to casual ones. Therefore, one of the keys to successfully involve a large population is to lessen the burden of the task. From this point of view, a minimum unit of the task participation in the framework is introduced as one very short segment of speech, which is called here as "speech chunk", so that a person can join the task if only he or she has time of transcribing, for instance, a few minutes. Speech chunk is a short speech block that contains a few phrases or sentences of averagely 3 or 4 seconds. Speech chunks are extracted from audio content data by a speech chunking front-end, in which a tailored speech activity detection is applied for this purpose. It is incorporated in the transcription framework aiming at increasing the task allocation flexibility and reducing the task burden itself.

II. Transcription Framework

A. Outline

The basic outline of the speech transcription system is shown in Fig. 1. It consists of three parts, speech chunking front-end, speech chunk transcription, and text processing back-end. Prior to the speech transcription stage, the audio data of a given digital content is processed by the speech chunking front-end to extract all the speech chunks in the content. These speech chunks are then passed to the transcription stage for participants. The basic task operation by participants is quite simple. That is, to access the Web site from their client environments, listen to one speech chunk,

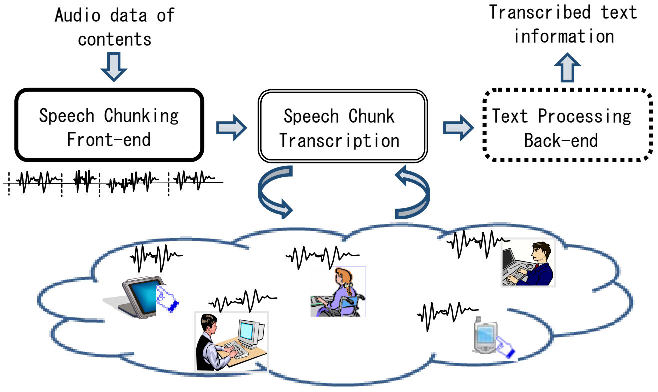
Audio data of contents
Transcribed text information
Speech Chunking Front-end
Speech Chunk Transcription
Text Processing Back-end

Fig. 1: Speech transcription framework

type in what is said, and finally submit it. Participant users may finish the task with only one chunk, or repeat it as they like. The transcribed texts from all the participants are collected together to the text processing back-end to build up a full transcription for a given content.

B. *Speech Chunk Transcription*

The work of participant users is only to listen to one speech chunk and type in what is said. In order for users to be able to easily catch the starting and ending portions of the current speech chunk for transcription, a "chunk block" that contains three consecutive speech chunks of preceding, current, and following ones is presented to users for playback. The time needed for transcribing one chunk would be at most a few minutes. Accordingly, in many cases, users might continue to work on another chunk, but it is not always the speech segment next in time to the current chunk. This is because each speech chunk is independently allocated to users. Therefore, they are assumed to work on the transcription by using only its local context. One advantage to this chunk-based transcription is that it can hide the whole story information from casual participants for a security purpose. On the other hand, global context might be helpful in some cases, for example, to disambiguate unclearly-pronounced phrases. In this framework, the disambiguation is to be done later in the text processing back-end or the transcription check in the final stage.

The user interface of speech chunk transcription is quite simple and no special functionality is required. By simplifying the interface, it can work not only on PCs, but also on various mobile devices as seen in Fig. 1. The system requests just one line input of text for a speech chunk. In case of small mobile devices on which text input is sometimes not so easy as PCs, speech recognition for parroting a phrase can be applied instead of typing it in.

C. *Text Processing Back-end*

The fragments of transcribed texts are collected together to the text processing back-end, and constructed to form the transcription data of a given content according to the time information attributes of speech chunks. For video contents, caption layout information is additionally generated from the time-stamped texts considering the maximum number of characters displayable on a screen.

Since our primary purpose of speech transcription is to make contents accessible for hearing impaired people, we are assuming here that possible participants are mainly volunteers on the internet. Although volunteers are expected to be a person of goodwill and their contributions are averagely expected as good quality, some sort of functionalities for checking and maintaining the transcription quality should be incorporated in the framework. From this point of view, information related with transcription quality such as statistics on text fluctuation or word confidence is also collected in this stage. Currently, these data are used for the final transcription check which is performed by a number of experienced users.

III. SPEECH CHUNKING FRONT-END

In [5], a kind of speech chunking method was applied to a micro task style of speech transcription by humans. The research work showed improvement in transcription quality by using the idea of Games With A Purpose (GWAP) [6]. As for the speech chunking, uniform 10-second segments were applied to divide a given content audio. The uniformly divided segments may not be appropriate in terms of listening, since the segment boundaries were generated irrelevantly to speech events such as onsets and end-points of speech or silence.

The speech chunking here is designed as a Voice Activity Detection (VAD) method for the purpose of extracting speech segments in an audio data stream. One typical application of VAD is for a preprocess stage of speech recognition systems (e.g., [7], [8]). The purpose of speech chunking here is somewhat different from that. Besides excluding unnecessary silences, it aims at extracting appropriate length of speech segments for the transcription task. Too long segments, for example, longer that 10 seconds, are hard to handle playback operations. On the contrary, too short segments less than 1 second make the transcription task inefficient because the overhead for setting up a task would not be negligible. For a feature for speech chunking VAD here, short-time speech power is used as a baseline, since power-based VAD methods generally perform well except for low SNR environment [8] that is not the case here. The algorithm of the speech chunking is shown in Fig. 2. First, candidates of non-speech intervals (NSI) are searched for as a low power region in the current audio speech interval (SI) to be processed. If the SI is longer than a threshold (Lc (the center threshold): assumed average length for speech chunks), then find the longest NSI candidate in the SI and divide the SI by the NSI candidate into two new SIs to be processed. Iterate the above steps until no more long SIs with NSI candidates are found. Each time in the iteration step, the signal power of the new speech interval is normalized to exclude the effect of power fluctuation in other intervals. In case where a too long SI (longer than a maximum length threshold: Lmax) which does not contain any low power region still exists after the basic iteration, forced division is applied by searching the power dip in there. As the result of the whole process, speech chunks around Lc in length are obtained as SIs.

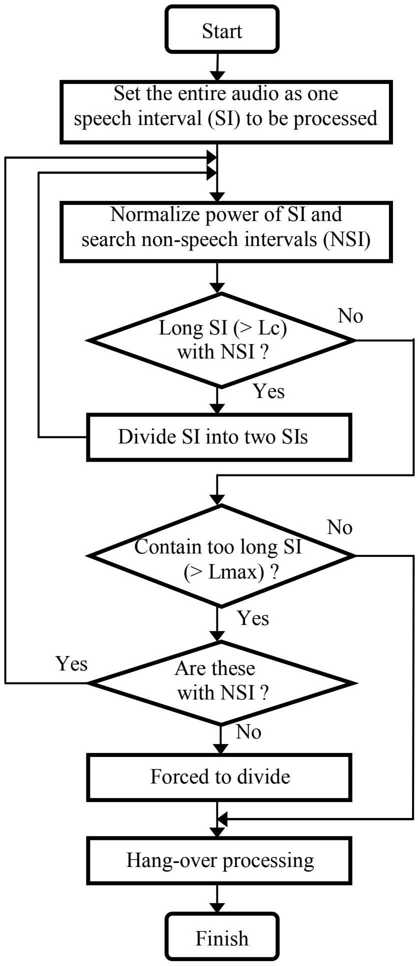
Start
Set the entire audio as one speech interval (SI) to be processed
Normalize power of SI and search non-speech intervals (NSI)
Long SI (> Lc) with NSI ?
No
Yes
Divide SI into two SIs
Contain too long SI (> Lmax) ?
No
Yes
Yes
Are these with NSI ?
No
Forced to divide
Hang-over processing
Finish

Fig. 2: Speech chunking algorithm

IV. Experiment

Here, the performance of the speech chunking front-end is investigated, which is the key component of the human-based speech transcription system. A quite broad range of digital contents in information, education, entertainment, etc. is expected to be handled for the purpose of content accessibility enhancement for hearing-impaired people.

A. Chunk Length Distribution

For experimental materials here, four different kinds of contents (news podcast, dialogue video, information channel podcast, animation video) are picked up as shown in Table I. The length of these content materials ranges from 10 to 20 minutes. In terms of constituent elements in audio data that might affect the chunking performance, these contents are classified by three factors: 1) number of speakers (Nsp), 2) with or without background music (BGM), 3) with or without sound effects (SE), as shown in Table I. The animation video is the most challenging content that includes many speakers, BGM, and dynamic sound effects.

In the algorithm of speech chunking shown in Fig. 2, Lc (the center threshold) and Lmax (the maximum threshold) are set as 4 and 10 [sec], respectively. These values were determined empirically through the algorithm design and prior experiments for more than 50 speech datasets. The distributions of speech chunk length obtained for all the contents are shown in Fig. 3.

TABLE I
CONTENT MATERIALS

Content ID	Category	Media	N_{sp}	BGM	SE
NW1	news	podcast	1	No	No
DG1	dialogue	video	2	No	No
IF1	information	podcast	2	Yes	No
AM1	animation	video	>10	Yes	Yes

(N_{sp}: number of speakers, SE: sound effects)

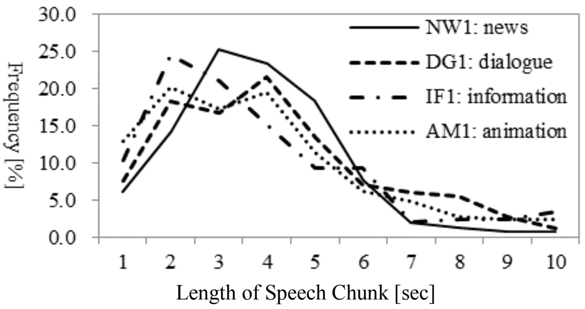

Fig. 3: Distributions of speech chunk length

As seen in the figure, most speech chunks are around 3 to 4 seconds in length regardless of content categories. The tendency on the whole follows what is intended for the speech chunking algorithm.

B. Chunk Type Analysis

This section discusses how appropriately speech chunks can be extracted as a task unit of short speech transcription, and the following points are to be checked: 1) whether a speech chunk does contain or not a linguistically meaningful segment (i.e., complete word or phrase segment) that can be transcribed, and 2) what sort of effects are posed on the transcription by various non-speech sounds. From this point of view, speech chunks are classified into four types as follows:

S (Speech): The chunk contains complete words or phrases, and it starts and ends with speech. It means that the transcribed text synchronizes in time with the chunk starting and ending positions.

M (Mixed): The chunk contains complete words or phrases, but it starts or ends with non-speech sound in this chunk. It means that the transcribed text may not synchronize in time with the chunk starting and end positions. It does not pose any problem for the transcription itself, but, when the transcription is used for video caption, this type of speech chunks may need a further adjustment for time synchronization.

N (Non-speech): The chunk contains only non-speech sounds such as BGM, or any other sound effects.

D (Divided): The chunk contains speech, but the speech includes a part of a word or phrase, which is divided apart into this chunk and adjacent ones.

Type S and M do not pose any problems in the transcription task at all. Type M, as stated above, may need time adjustment with motion pictures in case of transcription for video captions. Type N is not needed for speech transcription. When participants encounter this type of chunks, they have only to indicate that there is no text in it by clicking "no text" button. Although this operation is much easier than typing in text, it is desirable that it occurs less frequently. Type D includes a part of a word or phrase that needs adjacent chunks for transcription.

All the speech chunks obtained for the content materials are analyzed. First, the "false negative" error, which means that actual speech segments are NOT extracted as a speech chunk, is 0 % for all the materials. It is important in a sense that texts collected from speech chunks can ensure to generate the transcription for a given content. The occupancy rates of chunk types for each content material are shown in Table II. For the news podcast (NW1), 95% of the speech chunks are classified into type S, which brings no difficulty in the transcribing task. Chunks of type N (about 4%) contain noises arising from turning a page of news manuscript and a short music break inserted between topics. There is only one chunk of Type D, which contains a stop of saying to correct the phrase. As seen from the result, the speech chunking works quite well for news podcast.

Including the other contents of DG1 (dialogue video of two Japanese politicians), IF1 (information channel podcast) and AM1 (animation video), the tendency of the occupation rates of each speech chunk type is compared. As for type S, the occupancy rates are 88% (DG1), 75% (IF1), and 67% (AM1). These rates decrease in the order of DG1, IF1 and AM1. Even for the tough condition of animation video, more than 60% are extracted as type S chunks.

Next, for type D, occupancy rates are 9.7% (DG1), 21.6% (IF1), and 4.9% (AM1). There are number of phenomena observed for type D chunks. Typical examples are 1) stop of saying or rephrasing, 2) intra-word "Sokuon" (Japanese-specific double consonant), 3) prolonged ending of a phrase that may include silence. The occupation rate of type D in the information channel podcast (IF1) is much higher than any other contents. The main reason would be that the speaking style of IF1 is quite informal and conversational. Even for these type D chunks, the transcription does not pose a serious problem because a "chunk block" of three consecutive speech chunks mentioned in section II.*B*. is always presented to users, although the playback operations might be needed slightly more than for other types.

TABLE II
TYPE OF SPEECH CHUNKS (SC)

Content ID (# of SCs)	Type of speech chunks			
	Speech(S)	Mixed(M)	Non-sp. (N)	Divided(D)
NW1 (260)	248(95.4%)	0 (0%)	11 (4.2%)	1 (0.4%)
	835 [sec]	0 [sec]	24 [sec]	2 [sec]
DG1 (185)	164(88.6%)	0 (%)	3 (1.6%)	18 (9.7%)
	601 [sec]	0 [sec]	2 [sec]	71 [sec]
IF1 (204)	153 (75%)	3 (1.5%)	4 (2.0%)	44 (21.6%)
	509 [sec]	17 [sec]	8 [sec]	130 [sec]
AM1 (370)	248(67.0%)	35 (9.5%)	69 (18.6%)	18 (4.9%)
	840 [sec]	196 [sec]	147 [sec]	51 [sec]

(upper: #of SCs, lower: total time length)

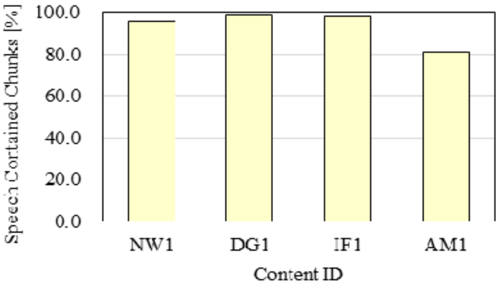

Fig. 4: Efficiency of speech chunking

The rate of chunks that contain actual speech segments over all the extracted chunks is shown in Fig. 4. The rate means how effectively the speech chunking front-end extracts speech utterances from the audio content data. The figure shows that, even for AM1 (animation video), over 80 % of speech chunks are extracted appropriately as a transcription task unit.

V. CONCLUSION

The primary advantage of the presented speech transcription framework is to reduce the burden of the task by chopping audio in advance into appropriate length of utterances and accordingly easing the task operation itself. The speech chunking front-end developed for this purpose was investigated for various types of contents in terms of several factors relevant to the chunk extraction performance. The result showed that it can be applied even to animation video contents that usually include dynamic sound effects. For further improvement, automatic detection of non-speech chunks (type N) is to be investigated to decrease the "no-text" check operations.

REFERENCES

[1] Munteanu, C., Baecker, R. and Penn, G., "Collaborative Editing for Improved Usefulness and Usability of Transcript-Enhanced Webcasts", Proceedings of ACM CHI2008, pp. 373-382, 2008.

[2] Goto, M. and Ogata, J. "PodCastle: Recent Advances of A Spoken Document Retrieval Service Improved by Anonymous User Contributions", Proceedings of Interspeech, pp. 3073-3076, 2011.

[3] Lasecki, W., Miller, C., Sadilek, A., Abumoussa, A., Borrello, D., Kushalnagar, R. and Bigham, J., "Real-Time Captioning by Groups of Non-Experts", Proceedings of the 25th annual ACM symposium on User interface software and technology, pp. 23-34, 2012.

[4] Parent, G. and Eskenazi, M., "Speaking to the Crowd: looking at past achievements in using crowdsourcing for speech and predicting future challenges", Proceedings of Interspeech, pp. 3037-3040, 2011.

[5] Liem, B., Zhang, H. and Yiling, C., "An Iterative Dual Pathway Structure for Speech-to-Text Transcription", Proceedings of Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence, pp. 37-42, 2011.

[6] von Ahn, L. "Games with a purpose", IEEE Computer Magazine 39(6), pp. 92-94, 2006.

[7] Karray, L. and Martin, A., "Towards improving speech detection robustness for speech recognition in adverse conditions", Speech Communication, 40(3), pp. 261-276, 2003.

[8] Chuangsuwanich, E. and Glass, J., "Robust Voice Activity Detector for Real World Applications Using Harmonicity and Modulation frequency", Proceedings of Interspeech, pp. 2645-2648, 2011.