PLAYING SPEECH BACKWARDS FOR CLASSIFICATION TASKS

Wolfgang Hürst, Tobias Lauer, Cédric Bürfent

Institut für Informatik, Albert-Ludwigs-Universität Freiburg, Germany
{huerst, lauer, buerfent}@informatik.uni-freiburg.de

ABSTRACT

Literature on speech skimming reports on techniques for playing speech backwards in a way that is still intelligible to the user. However, so far there is no empirical evidence for reasonable parameter settings of the respective algorithms and few examinations have been conducted to verify the usefulness of this feature for actual tasks. We present a user study testing different ways of backward skimming in relation to topic classification. Our evaluation shows a high classification performance and suggests implications for the design of the user interface.

1. INTRODUCTION

When skimming printed text, people generally do not follow the linear flow of the respective document: they read diagonally, skip passages, go back and forth, etc. Common techniques for speech skimming try to simulate this behavior by breaking the strict temporal characteristics of this media type, e.g., by replaying speech faster, allowing jumps to sentence borders, etc. Several systems also support some sort of backward replay which is realized in a way that leaves the content intelligible to some extent [2,4,6]: rather than playing speech backwards sample by sample, small segments of speech are played normally (i.e., in forward direction), but in reverse order (cf. Figure 1A). Different values for the length of the segments have been proposed in the literature. For example, [1] recommends lengths between 0.25 and 2 sec, while [6,4] use a segment length of 4 sec. Interestingly, none of the previous works provide an empirical basis for their choice of a specific value, nor is there more than some anecdotal information about usability or user performance when backward replay is applied to specific tasks

An obvious application of backward speech replay is topic classification in the context of skimming and searching speech. When skimming visual data such as text or video, it is natural for users to go back and forth in the document in order to locate specific information. Especially when skimming continuous, time-dependent data (e.g., video) at a higher speed, situations are very likely where users need to go back in order to verify whether or not the part just seen is indeed relevant. Therefore, backward audio seems to be a reasonable interface extension in similar situations when skimming audio files. The purpose of this paper is to verify this claim by examining the usefulness of backward speech replay for topic classification. Our main goal is to find out if there is any actual value in backward replay of speech signals and to analyze different parameter settings on how they influence users' perception and their overall performance in a classification task.

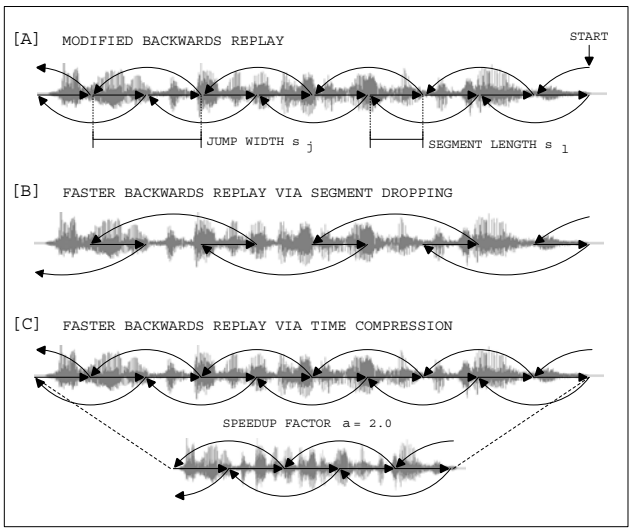

Figure 1. Approaches to backward replay.

2. APPROACHES TO BACKWARD REPLAY

Intelligible backward replay of speech signals can be achieved by continuously playing short snippets of audio in reverse order. The main parameters influencing the quality of the perceived signal are the segment length s_l of the single snippets and the jump width s_j (cf. Figure 1A). Increasing s_j to a value larger than $2 \cdot s_l$ enables users to skim the speech signal in less time because parts of the signal are dropped, as illustrated in Figure 1B. Although this results in a loss of information, this approach can be useful for tasks such as classification, as long as s_l is large enough to allow intelligible audio feedback. The situation can be compared to backward skimming of a printed text, where it is not necessary to read every word in order to get

0-7803-9332-5/05/$20.00 ©2005 IEEE

an idea of the overall content. Another approach for faster backward skimming of speech signals is to increase replay speed using time-scaling [5] (see Figure 1C). This approach has the advantage that, instead of larger chunks, only redundant information is left out. On the other hand, time-compressed speech can become harder to understand, even if played forward. [6] report that they use faster replay for backward skimming. However, no comments on the maximum speed value, the users' perception, or the overall usability are made.

3. EVALUATION

Setup. In the following, we present three experiments with the approaches for backward replay depicted in Figure 1A-C. For the first one, *experiment A*, we used standard backward skimming, as shown in Figure 1A. The segment length s_l was chosen to be the independent variable while the jump width s_j was set to be $2 \cdot s_l$. The speech data used consisted of news clips of 8 to 10 sec in length (extracts from radio news messages) and the task was to identify the topic and content of the corresponding news message.

Experiment A was subdivided into two tests. In the first one, users listened to one clip several times with different values assigned to the segment length s_l (in ascending order) and were asked to rate each value on a given scale. While literature reports the usage of rather long segment lengths of up to 4 sec (cf. Section 1), we believe that backward replay makes more sense if shorter segments are used: if segments become too long, users will not perceive it as playing backwards any more and will tend to use other mechanisms where they are in control of the jump positions. Based on initial testing we assumed 2 seconds to be a reasonable value for backward replay. Hence, the following segment lengths s_l were evaluated: 0.25, 0.5, 0.75, 1, 1.5, 2, 2.5, and 3 sec. Each participant had to listen to a speech file in reverse order with the particular segment length, and give a subjective rating based on the value, quality, and usefulness of this kind of backward replay. Ratings were done on a 5-point scale describing the parameter s_l as *"far too small"*, *"too small"*, *"ok"*, *"too large"*, or *"far too large"*. Users were allowed to listen to each version only once and subsequently had to make their relevance judgment. Later changes on the judgments were allowed (but no re-listening). The test started with the smallest value, where we expected comprehension to be very low, and the segment length was then increased according to the above values. Different speech files were used for different users, but each subject got the same file for all parameter values during this test. By this procedure, we hoped to identify a threshold value for the segment length at which replay becomes intelligible. (Note that a possible learning effect could not be excluded in this part. However, since the effect would apply to all test subjects in the same way, it was accepted since the goal was to find out whether or not such a threshold exists at all.)

The goal of the second test was again to evaluate different parameter settings, but in addition and more important, to see how users perform in a classification task where they have to classify different news clips according to their topic. The same set of parameter values was used as before, but this time in random order. The participants had to listen to a news clip backwards and then classify its content based on a given list of news topics. 20 clips were taken from a pool of 10 news messages, all of which were about sports, and used in experiment A to C. There were 3 messages about soccer, 3 about car racing, 2 about cycling, and 2 about other sports. Classification could be made for the actual news message (e.g., *"Schumacher injured in accident"*), the overall topic (e.g., *"car racing"*), or could be left out if users were not able to classify the clip at all. In addition, they had to give a subjective rating, as in the first test. This time, different clips were used for each parameter value. The mapping of a clip to a parameter value was equally distributed among the users. Hence, each user heard the same clips as the other participants, but in a different order and with different parameter values. Users were allowed to listen to each file only once and had to answer the questions immediately. No re-listening or later modification of the judgments was allowed in this test.

Users were encouraged to make comments during the tests – according to the common *think-aloud technique* for UI evaluation. After finishing both tests of experiment A, users had to perform experiments B and C. To exclude learning effects due to the order of the two experiments, half of the subjects started with experiment B, the others with experiment C.

Experiment B provided faster skimming by segment dropping, as depicted in Figure 1B, with the jump width s_j as the parameter to be evaluated. Based on our initial testing, the segment length s_l was set to a fixed value of 2 sec for this test. Again, users had to perform two tests which were set up in the same way as experiment A. However, in this case we started with the best possible value, i.e. $s_j = 4$ sec and subsequently increased the jump width s_j. The following values were evaluated: 4, 4.5, 5, 5.5, 6 and 6.5 sec (ascending order for test 1, random order for test 2). User ratings were again based on a 5-point scale, this time ranging from *"very good"* to *"very bad"*.

Experiment C was set up in the same way as experiment B, but this time using faster replay as illustrated in Figure 1C. Hence, the speedup factor α served as the independent variable and the following values were used: 1, 1.25, 1.5, 1.75, 2 and 2.25 times normal replay speed. It should be noted that these values were chosen in order to achieve the same overall time compression rates in both cases, experiment B and C. Again, a fixed segment length s_l of 2 sec was used.

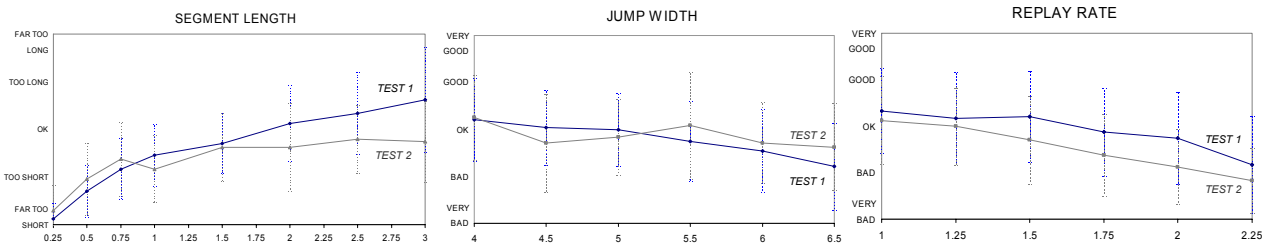

Figure 2. Subjective user judgements for both tests of experiment A, B, and C, respectively.

After completing all 3 experiments, each participant was interviewed and had to answer a short questionnaire. The overall evaluation time per user was between 20 and 30 minutes. 24 users participated in the evaluation: 13 male, 11 female, aged 16 to 61 (with 14 users between 20 and 29). Eight of them were students. All others had different professional backgrounds. None of them had any experience with backward speech replay.

Results. Although some users were very skeptical about the usefulness and feasibility of backward replay before the experiments, 18 out of 24 agreed afterwards that it is a useful feature for speech skimming. In addition, 23 users thought backward audio replay could be a very useful enhancement for skimming and searching in audio-visual documents, e.g. for tasks related to video browsing. When asked about faster replay, 63% preferred time-compressed audio, while 29% preferred the method from Figure 1B. However, these personal preferences did not have any influence on the classification task, where all users performed equally well.

The results of the subjective user judgments (averaged over all participants) for both tests of each experiment are illustrated in Figure 2. The standard deviation indicates that there was a rather large variance in the answers among the users. When comparing the outcomes of the two tests with each other it is important to keep in mind that test 2 was always done after test 1, which means that users were more familiar with backward replay. However, in the second tests the order of the parameter values was randomized to eliminate learning effects, and the users did not just listen to the file but had to solve an actual task. Therefore, the subjective judgments were expected to be less regularly distributed and generally a little lower than in the first tests, an assumption that is confirmed by the data.

The rather large deviations in the subjective user judgments were a little surprising to us. A closer look at the data showed that judgments varied a) within one document between different users, and b) between documents for one single user. This indicates that perception of backward speech highly depends on the users' personal preferences as well as on the actual document. Hence, although most users considered backward replay to be useful, there was large disagreement on the best realization of this feature.

The results of the tasks in the second test of each experiment are illustrated in Figure 3. Users performed surprisingly well. Even parameter values considered "difficult" yielded predominantly correct results. For example, with a jump width of 6 sec (i.e., 2 seconds of every 4-second block were skipped), two thirds of the users were still able to identify the corresponding news message and another 17% were able to at least classify the overall topic. Performance with time-compressed audio decreased more obviously with higher values, but even at the highest rate of 2.25 times normal speed (a speedup rate at which normal forward replay usually starts to get incomprehensible), almost half of the users were able to identify the corresponding news message. In addition, it is interesting to note that all users performed equally well. For example, in experiments B and C no user made more than one mistake for the parameter values below 6 sec and a speedup rate of 2, respectively. Although some users had a clear preference for one of the two approaches, there was no difference in their performance in the classification task. In addition, no differences could be observed between different groups of users (male vs. female, age, etc.).

Considering the individual parameters, no clear trend could be identified, except for the obvious findings that shorter segments, larger jump widths, and faster replay obviously lead to a decrease in comprehension. The observation that a segment length of 750 ms yielded better classification results than 1 sec seems a little surprising. One possible explanation is that the typical speech rhythm and speed of the radio news used in this study might fit one value better than the other, i.e., cutting messages randomly into 750 ms portions may distort it to a lesser degree than using 1 sec pieces. However, further investigation is required to verify this observation. Against our expectations, the first experiment did not yield a clear threshold or a good value for the segment length. However, it became apparent that rather short segments are sufficient for classification. Even at the lowest value of 0.25 sec (subjectively rated as "far too short"), more than half of the users were still able to identify the overall topic correctly. Regarding faster

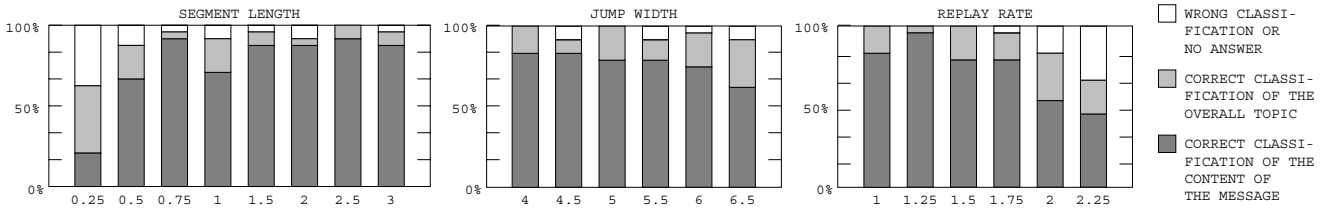

Figure 3. Results of the classification task (test 2) for experiment A, B, and C, respectively.

backward skimming, the variant which omitted parts of the signal (experiment B) showed a slightly better performance than the version using time-compressed audio (experiment C), although 63% of the users expressed a preference for the latter one.

4. CONCLUSION AND FUTURE WORK

The two main goals of our experiments were a) to analyze if backwards replay of speech signal is a useful technique in relation to classification tasks and b) to evaluate different approaches and parameter settings in order to be able to develop better tools and interfaces.

Considering the first goal, our evaluation proved that backward replay of speech signals can be a useful feature for topic classification, which is an important task when searching and browsing speech documents. This claim is based on the comments made by the test participants as well as on their performance in solving the tasks, which was much better than we expected. Although the classification task we evaluated in this study was rather easy (due to the fact that the users just had to pick one out of a fixed set of pre-given topics) it resembles a realistic situation, i.e. a user looking for news messages of particular events. More challenging classification tasks where users, for example, have to identify a topic without any pre-knowledge of the data are part of our future work.

One important observation was that while the influence of specific parameter settings on the classification performance was negligible, users' perception varied strongly, depending on their personal preferences as well as on the actual documents. While most interfaces for backward replay of speech only offer very limited possibilities and restricted freedom for manipulation of the involved parameters, there does not seem to be one "best" solution for this case. Based on this finding, we conclude that user interfaces for backward skimming should not be restricted to a single form and parameter setting. Instead, they should offer flexibility not only for forward skimming (where it is common for users to be able to choose between different parameter settings, such as different replay speeds), but also for backward skimming. However, it is important that the interface does not get too complex and overloads users with features they are unlikely to use. Advanced user interfaces for searching and skimming speech such as the one proposed in our previous work [3] can be further enhanced by integrating backward replay with the existing features. For example, combining backward audio with real-time interactive manipulation of replay speed would allow users to skim speech files in both directions at flexible replay rates using one easy-to-learn interface.

A further issue for future work is the analysis of an adaptive, more "intelligent" choice of the segment length based on automatic detection of sentence, word, and sub-word boundaries. Although our first experiments in this direction did not show any real improvement, there is some evidence in the data indicating that most of the incorrect classifications in the second tests of experiments A and B were caused by the fact that many backward jumps ended up in the middle of some important words.

5. REFERENCES

[1] B. Arons, "Authoring and transcription tools for speech-based hypermedia systems," *Proceedings of AVIOS 1991*, 15-20, 1991.

[2] B. Arons, "SpeechSkimmer: A system for interactively skimming recorded speech," *ACM ToCHI 4* (2), 3-38, 1997.

[3] W. Hürst, T. Lauer, and G. Götz, "Interactive manipulation of replay speed while listening to speech recordings," *Proceedings of ACM Multimedia 2004*, 488-491, 2004.

[4] J.S. Kim, "TattleTrail: An archiving voice chat system for mobile users over Internet Protocol," Masters thesis, Massachusetts Institute of Technology, 2002.

[5] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," *Proc. ICASSP 85*, 493-496, 1985.

[6] C. Schmandt, J.S. Kim, K. Lee, G. Vallejo, and M. Ackerman, "Mediated voice communication via mobile IP," *Proceedings of UIST 2002*, 141-150, 2002.

Acknowledgments: This work was supported by the German Research Foundation (DFG) as part of its research initiatives "Distributed processing and exchange of digital documents (V3D2)" and "Net-based knowledge communication in groups".