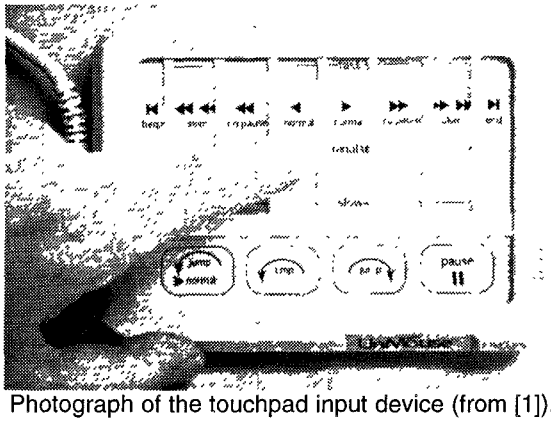
Hands-On Demonstration: Interacting with SpeechSkimmer

*Barry Arons**
Speech Interaction Research
PO Box 14
Cambridge MA, 02142-0001
E-Mail: barons@media.mit.edu
Phone: +1 500-DoSound

ABSTRACT
SpeechSkimmer is an interactive system for quickly browsing and finding information in speech recordings. Skimming speech recordings is much more difficult than visually scanning images, text, or video because of the slow, linear, temporal nature of the audio channel. The SpeechSkimmer system uses a combination of (1) time compression and pause removal, (2) automatically finding segments that summarize a recording, and (3) interaction techniques, to enable a speech recording to be heard quickly and at several levels of detail.

SpeechSkimmer was first presented at UIST '93 [1]. Since that time several important features have been added (see [2]). Most notable is the use of a pitch-based emphasis detection algorithm to automatically find topic introductions and summarizing statements from a recording [3, 4]. This demonstration is presented as a hands-on guide, allowing one to explore the SpeechSkimmer user interface.

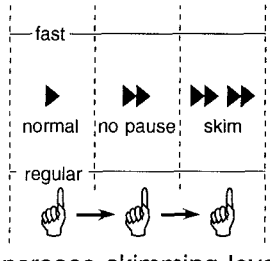
KEYWORDS
Speech skimming, time compression, non-speech audio.



Photograph of the touchpad input device (from [1]).

TO DO AND NOTICE
Start the demonstration by going to the beginning of the recording by touching in the |◄ area at the left of the touchpad. Listen to the different skimming levels and time compressions as described in the following sections. The finger pointer (☜) in the figures indicates where to touch.

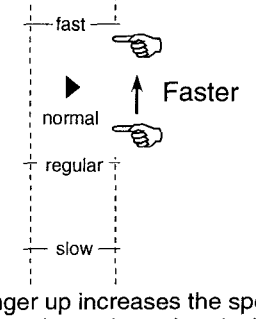
SKIMMING LEVEL 1: NORMAL
There are three skimming levels: normal, pauses removed, and automatically skimming selected segments. Play a sample of level 1 speech (touch ►). Listen to the talker's natural speaking rate, and how the talker uses pauses and intonation for emphasis.



Increase skimming level

Moving your finger horizontally changes the skimming level.

TIME COMPRESSION
A simple time compression algorithm ("sampling" with a crossfade) is used to change the playback speed of the recording without changing the voice characteristics. A range of time compressions from regular speed (1.0x) to 2.2x is available on this PowerBook 170 demonstration. Recordings can also be slowed down to 0.6x normal (using a different time scaling algorithm).



Moving your finger up increases the speed of playback, moving down slows the playback.

When listening with headphones, notice that as the speed is increased above 1.0x the signal is presented dichotically (i.e., the speech information is split between the ears). With this technique no information is discarded until the speed is above 2.0x. Most people find the dichotic presenta-

*This work was performed at the MIT Media Laboratory, and was sponsored by Apple® Computer, Inc. and Interval Research Corp.

UIST 95 Pittsburgh PA USA
© 1995 ACM 0-89791-709-x/95/11..$3.50

tion over headphones easier to listen to, and more intelligible than, a single channel of time compressed speech over loud speakers or headphones.

The amount of time compression can be changed within any skimming level. Note, however, that there is a slight delay between when you move your finger and when the change in time compression takes effect.

LEVEL 2: PAUSE REMOVAL

Play level 2 speech (touch in the ►► region at regular speed). Notice that speech rate is faster (pause removal is a form of time compression), but that some of the words may sound slightly clipped at the beginning or end.

Play the level 2 speech with maximum time compression. Notice that combining pause removal with time compression can make the recording difficult to listen to as the listener is robbed of cognitive processing time.

LEVEL 3: SKIMMING BASED ON PITCH

Playing level 3 (►► ►►) gives a summary of the recording. Short segments of speech (~5 sec) are played with a brief silence between segments.

Notice that many of the segments are clearly emphasized by the talker, or are introductions to new topics. However, the algorithm sometimes selects segments that are not good summary statements (see section below on Jumping). Also notice that it becomes difficult to listen to the skimmed speech if you increase the playback speed, as the short isolated segments do not provide enough context to allow you to follow the monologue.

If you are skimming or playing at high speed, you can use the key on the bottom left to take you back a bit (↶), and play normally (►) at 1.0x.

PLAYING BACKWARD

Along with playing or skimming *forward*, you can also play *backward* (◄ or ◄◄) or skim *backward* (◄◄ ◄◄) allowing you to find a particular word or phrase that you just passed. When going backwards, short segments (selected based on pauses in the speech) are played in reverse order.

Listen to about 10 seconds of the recording by playing forward at regular speed (with pauses removed ►►). Now play the same passage backward (touch ◄◄ at regular speed). Notice that the overall meaning is incomprehensible since the word order is jumbled, but that the individual words and phrases are intelligible.

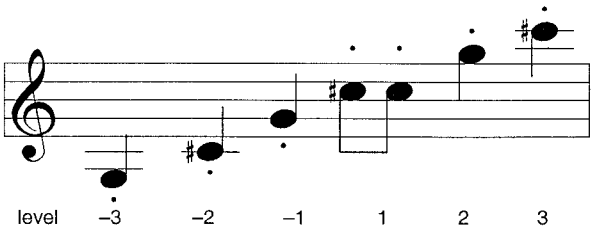
JUMPING

Besides using the various skimming levels to listen to a recording, you can interactively jump forward or backward using the "keys" at the bottom of the touchpad. This is useful if you are listening for a particular idea or topic, allowing the listener to jump quickly to the area of interest.

Begin playing at whatever skimming level and speed is comfortable. Touch the jump forward key (↷) or the jump backward key (↶) once or several times. Note that when in level 3 skimming (►► ►►) you are jumped to the next emphasized segment.

NON-SPEECH AUDIO FEEDBACK

Notice that when you change skimming levels a short tone is played; the frequency of the tone indicates the new skimming level. Try sliding your finger across all six of the skimming levels and notice the scale that is played.



A musical representation of the tones played at the different skimming levels. Notice the double beep "landmark" for normal (level 1) playing. The small dots indicate staccato (i.e., short and crisp) notes.

Notice the sounds that are played when you hit the end of the recording (you can go to the beginning of the recording with |◄, or the end with ►|). Also note that different sounds are played when you jump forward versus backward.

SKIMMING WITH SPEAKER ID

In addition to the pause-based skimming, SpeechSkimmer can be used to quickly browse through a dialogue. By using a speaker identification algorithm [5], level 1 (►) can be used to hear the entire conversation, level 2 (►►) can be used to hear one side of the dialogue, and level 3 (►► ►►) can be used to hear the other side.

CONCLUSIONS

This brief interactive exploration of the SpeechSkimmer system has demonstrated some of the techniques that can be used for quickly browsing speech recordings. These and related techniques can be incorporated into a wide variety of applications allowing speech to be used easily and efficiently in new user interfaces.

REFERENCES

1. Arons, B. SpeechSkimmer: Interactively Skimming Recorded Speech. In *Proc. of UIST '93.* ACM Press, Nov. 1993, 187–196.
2. Arons, B. "Interactively Skimming Recorded Speech." Ph.D. dissertation, MIT, 1994.
3. Arons, B. Pitch-Based Emphasis Detection for Segmenting Speech Recordings. In *Proc. of Intl. Conf. on Spoken Language Processing (Yokohama, Japan, Sep. 18–22),* vol. 4, 1994, 1931–1934.
4. Chen, F.R. and Withgott, M. The Use of Emphasis to Automatically Summarize Spoken Discourse. In *Proc. of the Intl. Conf. on Acoustics, Speech, and Signal Processing,* IEEE, 1992, 229–233.
5. Reynolds, D.A. and Rose, R.C. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models." *IEEE Transactions on Speech and Audio Processing 3,* 1 (1995): 72–83.