# SESLA TRANSCRIBER: A SPEECH TRANSCRIPTION TOOL THAT ADAPTS TO YOUR SKILL AND TIME BUDGET

*Matthias Sperber*[1]*, Graham Neubig*[2]*, Satoshi Nakamura*[2]*, Alex Waibel*[1]

[1]Karlsruhe Institute of Technology, Institute of Anthropomatics and Robotics, Germany
[2]Nara Institute of Science and Technology, Augmented Human Communications Laboratory, Japan

## ABSTRACT

We present a speech transcription tool targeted at situations in which cost is a critical or limiting factor. This tool actively guides the transcription process by taking an automatically created transcript as a starting point, and asking for correction of only the parts likely to contain errors. The transcriber specifies a time budget, and the software uses models of transcription accuracy and cost to choose which segments should be transcribed to achieve the highest error reduction. This approach has been found to be 25% more efficient than cost-insensitive approaches in previous work. The cost model is adapted to the transcriber on-the-fly during the transcription process, so no user enrollment is necessary. The segmentation is updated regularly to reflect improved cost models, and to recover from potential time prediction errors. The user interface was designed to be easy to learn and efficient to use. It allows either transcribing each segment from scratch or post-editing, and has logging features that allow detailed user studies.

## 1. COST-SENSITIVE TRANSCRIPTION

This paper describes a new tool for efficient manual correction of speech transcripts. Our tool uses an automatically created transcript as a starting point, and guides the transcriber through the correction of a selection of segments that are likely to contain errors. The general strategy is to focus only on those erroneous parts, and trust the speech recognizer for other parts, in order to cut transcription costs.

Our tool asks the user to specify a time-budget (for example, 30 minutes of annotation), and automatically chooses an appropriate number of segments for correction such that the time-budget is kept. The locations and sizes of these segments are chosen such that the expected reduction of errors is maximized given the time budget, according to the SESLA method (Segmentation for Efficient Supervised Language Annotation) as described in [1]. Specifically, the tool predicts both transcription time and error reduction for transcribing any possible segment in a speech. Using these predictions, an optimal segmentation into segments to transcribe and segments to skip is computed. There is a trade-off to consider when choosing between smaller and longer segments: While choosing very small segments (e.g. single words) would allow the transcriber to really concentrate only on parts that have a very high probability of error, longer segments are desirable from a cognitive point of view as they reduce cognitive overhead due to context switches. A global constrained optimization strategy that is based on the time and error reduction predictions and considers all possible segmentations allows finding an optimal trade-off in a principled way. Savings in human effort of 25% were observed, compared to the traditional approach of choosing low-confident segments from a fixed segmentation.

To obtain the transcription time and error reduction predictions needed to find the optimal segmentation, the tool proceeds as follows. Confidence scores provided by the automatic speech recognizer are employed to estimate chance of error. Transcription time is predicted via Gaussian Process regression [2], with the features segment length, audio duration, and average word confidence. The model is continually retrained on the observed transcription times during the ongoing transcription process. The regressor is initiated with a sensible prior so that rough predictions are possible even for new users, no initial user enrollment is required. Based on the prior model, the regressor is then retrained regularly to reflect the transcriber's characteristics with gradually increasing accuracy. To take advantage of the improving user models, the segmentation of the remainder of the speech being transcribed is updated regularly as well. Each segmentation update reflects the updated currently remaining time budget, and in this way, for instance, allows removing (skipping) less promising segments if the transcriber's progress has been slower than expected. Hence, we can rapidly recover from prediction inaccuracies that are expected to occur during practical use, and make sure that the remaining time is used optimally. These updating strategies have been proposed in [3], along with a fast segmentation algorithm which we employ in our tool.

## 2. USER INTERFACE

We took care to design the user interface in a way that is both easy to learn, and allows fast and convenient operation for advanced users. It has evolved over several iterations based

on pilot studies with users of various backgrounds, including computer scientists, linguists, and low-cost transcribers with low-profile educational background. To use our tool, a speech file along with a confidence-annotated automatic transcript must be provided. It is intended for correcting speech files of medium to long duration, such as lectures or meetings. First, a transcription time budget must be specified, and a trade-off curve with the expected number of errors that can be removed for different time budgets is displayed to aid the decision for a time budget (Figure 1).

As depicted in Figure 2, the transcription view displays the speech as a vertical waveform, and the transcriber can click anywhere in the waveform to play back the audio from that position. The selected segments are aligned next to it, with segments to transcribe colored white and segments to skip colored in gray. Transcription occurs in linear order from top to bottom. The active (white) segments can be transcribed or corrected by simply clicking on that segment. The corresponding audio will be played automatically, and can be repeated if necessary. Transcription is supported both from scratch or post-editing style, depending on an initial setting. Segments that should not be corrected are always displayed with their automatic transcript. Clicking on them causes the corresponding audio to be played, but they cannot be edited. After finishing a segment, the transcriber can confirm the segment, or mark it as unsure. Advanced users have complete control over the user interface via keyboard shortcuts. The specification of the time budget is the only input needed to enable the cost-sensitive transcription; if no budget is set, the whole talk will be transcribed. User models and the segmentation are updated fully automatically in the background without disrupting the transcription workflow. Longer periods of inactivity are recognized, and not subtracted from the time budget. In addition, outlier segments, such as if the transcriber got distracted, are automatically recognized and not used for the user model training. Multiple users are supported, with separate user models being trained for each user, but the transcription of the same speech cannot be shared between several transcribers. The tool logs detailed user activities such as keystrokes and audio playback to help answer research questions.

Our transcription tool can be used on Windows and Macintosh computers, and is available for academic use from `http://msperber.com/sesla-transcriber`.

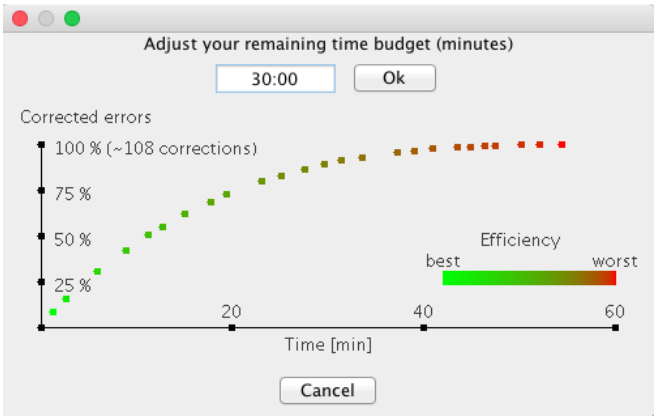


**Fig. 1**. Time budget prompt with visualized trade-off curve.

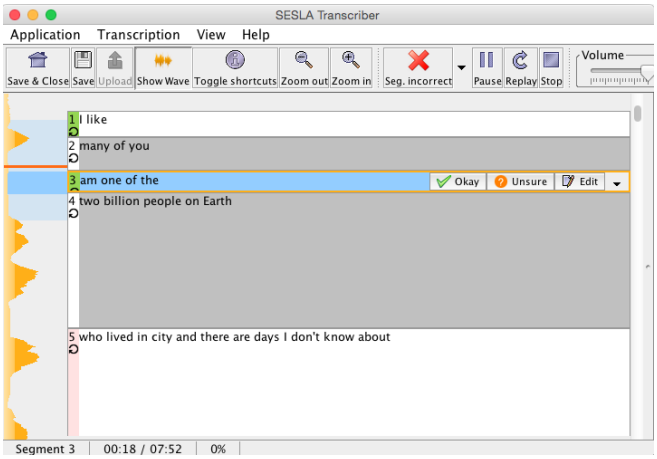


**Fig. 2**. The main transcription screen.

## Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 287658 Bridges Across the Language Divide (EU-BRIDGE).

## 3. REFERENCES

[1] Matthias Sperber, Mirjam Simantzik, Graham Neubig, Satoshi Nakamura, and Alex Waibel, "Segmentation for Efficient Supervised Language Annotation with an Explicit Cost-Utility Tradeoff," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 2, no. April, pp. 169–180, 2014.

[2] Carl E. Rasmussen and Christopher K.I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, USA, 2006.

[3] Matthias Sperber, Graham Neubig, Satoshi Nakamura, and Alex Waibel, "On-the-Fly User Modeling for Cost-Sensitive Correction of Speech Transcripts," in *Spoken Language Technology Workshop (SLT)*, 2014.