

Keith Vertanen

RESEARCH STATEMENT

While speech recognition technology has made amazing progress in the last few decades, it is still an imperfect technology. Recognition errors will occur and those errors can cripple the usefulness of the overall speech application. Whether the application is the dictation of email, the transcription of online videos, or the indexing of large audio archives, a successful speech application often requires provisions for human-guided error correction. My research focuses on how to make this human-guided correction process more efficient, less frustrating, and more accessible.

The primary principle I employ to improve the correction process is to “use all the information”. Speech recognizers work hard to search through a multitude of possibilities before deciding on the best recognition result. But when this best result is wrong, there is still useful information that can be gleaned from the recognizer’s search. In addition to the recognizer, the user is also a rich source of information. For example, we might utilize a user’s past corrective actions or monitor where the user is looking in the interface. By effectively combining all the information sources and exposing the various hypotheses to the user, I aim to reduce the user’s workload and make the correction process more seamless.

To demonstrate the utility of this principle, I developed two interfaces during my PhD: Speech Dasher [1] and Parakeet [2]. Speech Dasher uses a continuous zooming interface to allow navigation through the space of recognition hypotheses (represented by a recognition lattice). For words not in the lattice, Speech Dasher supports easy fallback to a letter-by-letter predictive spelling of words. In a user study, participants used speech and gaze to write at 40 corrected words per minute (wpm). This was despite a word error rate (WER) of 22%. The Parakeet interface is based on selecting words from a word confusion network. In computational experiments, about half of all errors could be corrected via Parakeet’s word confusion network interface. A predictive software keyboard allowed users to correct any error. In a user study, participants wrote at 13 wpm while walking outdoors. This was despite a recognition WER of 26% and significant recognition delays. Neglecting the delays, users could have written at up to 26 wpm. As a reference point, users of T9 wrote at 16 wpm while seated indoors after 15 sessions [3].

Looking ahead, I plan to continue exploring novel speech interface designs. In particular, I am interested in designs that push the boundaries with respect to feedback and input mechanisms (e.g. interfaces without visual feedback, voice-only input, etc). I also plan to expand into other recognition-based interfaces that use non-speech input sources (e.g. eye tracking and touch-screen gestures). Current projects include:

- One-step voice-only (or voice+eye) correction of recognition errors [4].
- Hands-free eyes-free text entry via speech.
- Dwell-free eye-typing using speech recognition inspired techniques.
- Comparing mobile text entry using speech and using an on-screen keyboard.
- Language model training in the email domain using large amounts of out-of-domain data.
- Crowdsourcing for speech and language data collection and for human factors experiments.

To summarize, my research hinges on the idea that interfaces based on recognition technologies need not have perfect recognition. Even at high error rates, interfaces that leverage all of the available information can be efficient and easy to use. A good correction interface will degrade gracefully, gleaning whatever useful information is available from the recognizer and from the user. In this manner I believe we can improve the productivity and satisfaction of users. By improving the user experience, speech applications will become more prevalent and impart greater utility to our world.

[1] K. Vertanen and D. J. C. MacKay. Speech Dasher: Fast Writing using Speech and Gaze. In *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, to appear.

[2] K. Vertanen and P. O. Kristensson. Parakeet: A continuous speech recognition system for mobile touch-screen devices. In *IUI '09: Proceedings of the 14th International Conference on Intelligent User Interfaces*, pages 237-246. ACM, 2009.

[3] J. O. Wobbrock, D. H. Chau, and B. A. Myers. An alternative to push, press, and tap-tap-tap: Gesturing on an isometric joystick for mobile phone text entry. In *CHI '07: Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 667-676. ACM, 2007.

[4] K. Vertanen and P.O. Kristensson. Automatic Selection of Recognition Errors by Respeaking the Intended Text. In *ASRU '09: IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 130-135. IEEE, 2009.