1aSCb1. Precision and error of automatic speech recognition.
Mark VanDam* and Noah H. Silbert

*Corresponding author's address: Speech and Hearing Sciences, Washington State University, Health Sciences Bldg, Room 125S, Spokane, WA 99202, mark.vandam@wsu.edu

Experienced judges assessed performance of an automatic speech recognition system developed for linguistic exchanges within families in their natural environment. Preliminary results suggest overall good performance with relatively high precision and low error.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

The value of automatic speech recognition (ASR) in speech communication research lies, in large part, in its overall ability to accurately label acoustic events that are relevant to language use. The automatic speech recognition software developed by the LENA Research Foundation (Boulder, CO) is an increasingly important tool in a variety of fields, including speech and language development and psycholinguistics. Using day-long recordings collected from a child-worn audio recorder, the software segments the audio and assigns labels to segments of speech. The labels include those for KEY CHILD, ADULT MALE, and ADULT FEMALE. Its successes and advantages notwithstanding, performance of the system has not been thoroughly reported.

Performance of the system is a serious concern for all ASR. Several reports in the literature have investigated reliability and accuracy of the LENA device, finding similar results across studies, each confirming reasonable levels of performance. In several reported studies, segments automatically labeled ADULT and KEY CHILD were similarly coded by human transcribers in 68% and 70% of instances, respectively [1-3]. Another study [4] reported reliability of automatic coding from a broader range of segments, concluding that there was about 64% agreement between automatic and human labels for the KEY CHILD category (the ADULT label was not considered). Additional studies have looked at reliability of the automatic algorithms to code specific acoustic parameters [4] as well as inter- and intra-judge reliability among human transcribers of acoustic recordings collected by the LENA [3-5]. Finally, a previous study looked at agreement rates between LENA labels and human judges, finding about 70% agreement [6].

Although these studies are indeed useful, additional evidence of the software's performance is necessary to better interpret and understand accumulating research using this tool. The objective of the present work is to investigate the LENA ASR software using different methodology from previous reports in order to assess performance of the system from a different perspective.

METHOD AND PRELIMINARY RESULTS

The present work assesses performance of LENA measures assessing live human vocal events; this study compares the labels assigned by the LENA ASR and those assigned by experienced human judges to the same acoustic signal. An important goal of the automatic labeling is to maintain relatively high precision by reducing false positives, especially with those segments that are labeled KEY CHILD. Another goal is to obtain high accuracy measures for all labels that are both highly sensitive (i.e., correctly assigned when present) and highly specific (i.e., correctly not assigned when absent).

The present study estimates precision between computer and human segment labels corresponding to children, mothers, and fathers, asking if certain label assignments have reliability estimates at different rates. For each of 26 families with typically developing children (age: $M=2.07$ yrs, $SD=0.69$ yrs), 30 segments of each talker machine-labeled as ADULT MALE, ADULT FEMALE, and KEY CHILD were pseudo-randomly selected and excised as audio clips from the daily recording. This yielded 2340 segments to be used as stimuli. Judges experienced in speech analysis then identified each segment as Mother, Father, Child, or Other, although there were no segments machine-labeled as Other. Judges' responses were analyzed in terms of agreement between human and machine as well as among judges and tested for precision and error rates.

Initial analyses of this on-going project indicate results that are broadly in accord with previous investigations, finding overall agreement between machine and human coding in the 70% range with Cohen's $\kappa > .55$. Also similar to previous findings, machine coding appears to be better at coding KEY CHILD segments than the ADULT segments, and agreements between machine and human judges were higher for ADULT MALE labels than for ADULT FEMALE labels. Further, the precision of ADULT MALE labels was similar to the rate for KEY CHILD segment labels, indicating that the ASR system performed similarly when assigning segment labels for children and fathers, but less well for mothers. Overall error rates were generally very low.

DISCUSSION

Overall, the findings in the present work are considered generally confirmatory, convergent evidence that the ASR methods used in the LENA software are precise. In particular, this work offers new evidence in terms of (1) more detail of talker identification expanding the previously reported category of ADULT to look at both mothers and fathers, (2) a larger number of judges, and, consequently, more exemplar tokens considered, and (3) a different sample of families and children from the data reported in previous literature. In conjunction with extant literature
looking at the performance of the system, this report offers further evidence of the validity of the LENA software. This research will further contribute to a better understanding of the limitations and advantages of ASR technology.

ACKNOWLEDGMENTS

Thanks go to Lisa Burton, Stephanie Nielsen, Aneesha Pretto, Jaimie Horning, Merry Spratford, Lauren Berry, Emily Sweet, Wendy Fick, and Rick Arenas for data management and work with families. This work is supported by the National Institute on Deafness and Other Communication Disorders and the National Institutes of Health (5R01-DC009560-01S1 to PIs Mary Pat Moeller and J. Bruce Tomblin). Content is the sole responsibility of the authors and does not necessarily represent the official views of the NIH/NIDCD.

REFERENCES


