VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

THE AUTOMATIC ANALYSIS OF CLASSROOM TALK

Hanna Kronholm[1], Daniela Caballero[2], André Mansikkaniemi[3], Roberto Araya[2], Sami Lehesvuori[1], Pasi Pertilä[4], Tuomas Virtanen[4], Mikko Kurimo[3] & Jouni Viiri[1]

[1]University of Jyväskylä, [2]University of Chile, [3]Aalto University, [4]Tampere University of Technology

ABSTRACT

The SMART SPEECH Project is a joint venture between three Finnish universities and a Chilean university. The aim is to develop a mobile application that can be used to record classroom talk and enable observations to be made of classroom interactions. We recorded Finnish and Chilean physics teachers' speech using both a conventional microphone/dictator setup and a microphone/mobile application setup. The recordings were analysed via automatic speech recognition (ASR). The average word error rate achieved for the Finnish teachers' speech was under 40%. The ASR approach also enabled us to determine the key topics discussed within the Finnish physics lessons under scrutiny. The results here were promising as the recognition accuracy rate was about 85% on average.

INTRODUCTION

When we are learning something, it is essential to interact with other individuals or cultural products–for example, books (Leach & Scott, 2003). Talk is one of the most common ways in which classroom interactions take place and, as such, it plays a significant role in learning. Where science teaching is concerned, the contents of the lessons are of importance because the adoption of scientific concepts is required (Driver, Asoko, Leach, Scott & Mortimer, 1994). Previous studies suggests that the richness and interconnectivity of the concepts of the lessons has a positive effect on learning (e.g. Helaakoski & Viiri, 2014). Thus, investigating what happens in the classroom, and what kind of interaction and talk is in the lessons is essential. Perhaps the most appropriate method for such research is classroom observation.

This project aims to develop an Android mobile application named SMART SPEECH, which will allow observations to be made using several different protocols, while recording teachers' speech via automatic speech recognition (ASR) at the same time. More specifically, we have selected three observation protocols, which give a general idea of the kinds of interactions that take place: The classroom observation protocol for undergraduate STEM (abbreviated 'COPUS') (Smith, Jones, Gilbert & Wieman, 2013), the communicative approach (Mortimer & Scott, 2003) and the ideas-objects relation (Millar & Abrahams, 2009). Here, COPUS is used for surface-level observation (i.e. what teachers and students do). The communicative approach is utilised to observe the interactions between a teacher and his/her students. Finally, the ideas-objects relation is for observing the links between the domain of objects and observable things, and the domain of ideas.

Existing studies of classroom talk tend to have involved manual analysis. This approach, however, is time-consuming and requires a high degree of precision. The development of automated analysis, therefore, would be a significant step in teaching and learning research. The speech-recognition approach should be seen as increasingly important for anyone working in the field of classroom talk. Indeed, a few studies on the automatic analysis of classroom talk have been carried out. Wang, Pan, Miller and Cortina (2014), for example, use the LENA™ system to classify classroom talk according to the speaker. In addition, Ranchal et al. (2013) convert teacher talk into text in real-time, while Blanchard et al. (2015) test the applicability of commercial speech-recognition systems to classroom-talk analysis. None of these existing systems, however, is able to convert speech into text and analyse the contents of the speech at the same time. In this study, we aim to develop a system that is capable of both.

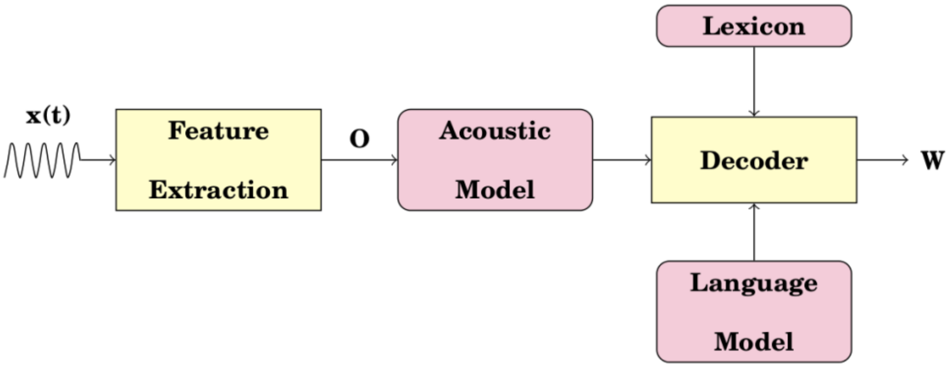

Figure 1. The architecture of a typical ASR system.

Recent advances in automatic speech recognition (ASR) mean that the performance level of such systems is potentially sufficient to allow large amounts of recorded classroom talk to be converted into text. A modern speech-recognition system includes the following components: a feature extractor, an acoustic model, a language model, a lexicon, and a decoder (Figure 1). Such ASR systems function as follows: Features, which carry important information about speech, are captured via an audio signal and then extracted from the recording. Based on the feature vectors, the acoustic model is used to estimate the probability of the occurrence of phonemes during minor intervals in the speech signal (Bahl & Jelinek, 1975; Juang et al., 1986). The language model is a statistical probability model for the occurrence of word sequences (Jelinek et al., 1975). The lexicon contains pronunciation rules (phoneme sequences) for all the words in the vocabulary. The decoder is the algorithm that finds the most probable word sequence that has been spoken, based on the speech signal and the statistical models (Young et al., 1989).

In recent years, progress in speech recognition has mainly been driven by the use of deep neural networks (DNNs) in acoustic modelling (Hinton et al., 2012). Feedforward and different variations of recurrent neural networks, such as long short-term memory (LSTM) cells, have replaced the traditional Gaussian mixture models in the estimation of phoneme probabilities.

Acoustic models should be trained on vast amounts of transcribed speech data and language models should be trained on large text corpora. The accuracy of an ASR system often depends on how well the testing data matches the training data. In the case of automatically transcribing STEM classroom recordings, one of the challenges is gathering enough topic-specific texts for training the language model. Where Finnish speech recognition is concerned, another challenge is related to the significant difference between colloquial and written language. There is often less colloquial text data available, which affects the recognition accuracy for conversational speech (Enarvi & Kurimo, 2013).

Speech recognition can enable the automatic analysis of classroom talk, even if the system being used produces a lot of erroneous output. If the error rate for topic-specific keywords is low enough, then different topic segments can be detected via a recording. ASR output is vectorised using aspects such as word weights (which indicate the importance of an individual word to the segment under scrutiny). Typically, word weights are calculated using the term frequency-inverse document frequency (tf-idf) measurement (Spärck Jones, 1972). The word-weight vector is often transformed into lower space using methods such as latent semantic indexing (LSI) (Deerwater et al., 1990) or latent Dirichlet allocation (LDA) (Blei et al., 2003). The idea behind these transformations is to tie together topic-related words into one common dimension. Topic segments can be detected by comparing the vectors generated via the ASR output.

In previous research, topic models have been used to detect topic changes in multiparty conversations (Sapru & Boulard, 2014) and lecture videos (Yamamoto et al., 2003), for example. In the present study, we implement a simple approach to test whether enough topic-related words can be recognised correctly to make automatic topic segmentation feasible. First, a set of topic classes is defined for each recording. Then, for each topic class, a list is composed consisting of the words that are related most closely to the topic. The segments in the recording are assigned to belong to a topic depending on how many topic words occur in them. Based on the ASR transcripts and the manual topic definitions, a classification score for each segment is calculated.

RESEARCH AIMS AND QUESTIONS

The aim of this paper is to describe the results achieved via the ASR system developed here. Our research questions are:

1. How well does ASR perform when interpreting lesson speech? In particular, what is the keyword error rate?
2. To what extent can ASR be used for the automatic recognition of lesson topics?

RESEARCH METHODS

This is an exploratory design study. We have obtained data (audio recordings) from Chilean and Finnish schools. As mentioned previously, the SMART SPEECH mobile application (Figure 2), which has been developed for this project, can be used for classroom observations and audio recordings. However, the ASR aspect of the study still has to be conducted via a computer. The mobile application has two functions: one through which the teacher can record his/her own lessons (teacher, in figure 2) and another via which classroom observations can be made (observer, in figure 2). It is not intended, however, that the application is used for both functions at the same time or that the teacher should observe his/her own lessons.

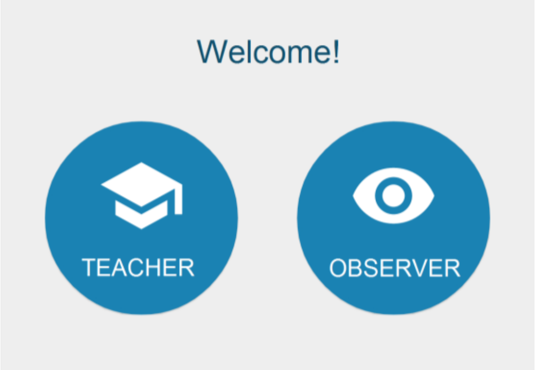

Figure 2: The SMART SPEECH mobile application.

In order to achieve better audio quality when recording teachers' speech, the application should be used with an external microphone. In the preliminary phases of the project, we used the SMART SPEECH mobile application with the RØDE Smartlav+ external microphone in both Finland and Chile. As the project has progressed, we have continued to utilise the smartphone to record in the Chilean classroom, but a ZOOM H4N dictator and two external microphones (AKG C250 and DPA SC4060-BM) have been used in the Finnish context because Finnish is challenging for ASR. When using an ASR system, there are benefits to having such high-quality audio recordings, especially at the development stage. At this point in time, therefore, the external equipment has enabled us to achieve more consistent and clear ASR results in the Finnish cases. The classroom is a unique and challenging environment, in which background noise can hamper data collection. Taking advantage of the dictator and microphone setup has enabled us to minimise the impact of such distractions.

To date, our recordings have involved teachers only. In the future, however, we would also like to record students' speech. At present, we have recorded four physics teachers in Finland. Altogether, we have recorded and analysed 17 lessons (of 30–45 minutes) in Finland. In Chile, we have recorded 45 physics and mathematics lessons (some of these have been double lessons of more than an hour, thus giving a total of 28 hours) and analysed two of the lessons thus far.

When running the Chilean ASR experiments in Spanish, the Google Cloud Speech API (https://cloud.google.com/speech/) was used. This tool supports over 80 languages, including both Latin-American Spanish and Finnish. In our early experiments, we gained higher levels of recognition for Finnish speech with the free, open-source Kaldi toolkit than with the Google Speech API. For this reason, the ASR experiments conducted in Finnish were run using the Kaldi toolkit (Povey et al., 2011).

Our recognition system was based on time-delay neural networks (TDNNs) combined with long short-term memory (LSTM) layers (Povey et al., 2016). The acoustic-model training data consists of speech from Speecon (Iskra et al., 2002) and the Finnish Parliament Speech Corpus. Altogether, we gathered 1708 hours of data from 782 speakers. An n-gram language model was trained on a subset from the Suomi24 corpus (http://urn.fi/urn:nbn:fi:lb-2017021506), which consisted of 76 million words of text retrieved from an online discussion forum.

The test data consists of 17 physics lessons recorded in Finnish classrooms between 2016 and 2017. The ASR results are reported via word error rate (WER) and keyword error rate (KER). *WER* is defined as follows:

$$WER = \frac{S + D + I}{N},$$

where *N* is the total number of words in the reference transcription, *S* is the number of substituted words, *D* is the number of deleted words and *I* is the number of inserted words.

In addition, *KER* is calculated as follows:

$$KER = \frac{S_k + D_k}{N_k},$$

where N_k is the number of keywords in the transcription, S_k is the number of keywords substituted and D_k is the number of keywords deleted. In this study, 'keywords' refers to terms that are related to physics.

Our second research question concerns lesson topics. Here, we recorded several lessons on the same topic in order to gather more data for training the ASR system. At first, we searched for all the keywords in the manually created transcripts, before placing these keywords in groups (keywords related to the same topic were put into the same group). Next, we divided the recordings into two-minute periods and searched for the keywords within these periods. From this data, we determined the topics that existed within these two-minute periods. As two minutes is a quite long period in the classroom environment, we found that keywords usually appeared from a number of different topics. The key topic for the period was determined, therefore, according to which keywords were most relevant within that period. We performed the same procedure for the transcripts produced via ASR and then compared them with the manually determined topics.

RESULTS

The speech-recognition results for the Finnish physics lessons are given in Table 1. The keywords were selected manually, based on the key topic of the lesson. The average WER was under 40% and the average KER was nearly the same. This indicates the difficulty of performing this task, even when using state-of-the-art recognition models.

Table 1. Speech-recognition results for Finnish teacher speech data (10 h 11 min). The results are reported via the word error rate (WER) and the keyword error rate (KER).

ASR system	WER [%]	KER [%]
Kaldi toolkit	36.3	33.9
Google Speech API	45.5	35.8

On average, better recognition accuracy was achieved with the open-source Kaldi toolkit than with the commercial, cloud-based ASR service. Due to the small amount of data analysed in the Chilean context, the results from the Finnish classroom only are reported in this paper.

We also noticed that there were differences in error rates between each of the teachers, as shown in Table 2. This reveals that the ASR system does not work as well for each and every speaker.

Table 2. Speech-recognition results for each Finnish teacher. The results are reported via the word error rate (WER) and the keyword error rate (KER).

	Kaldi toolkit		Google Speech API	
Teacher	WER [%]	KER [%]	WER [%]	KER [%]
A	44.7	42.9	47.6	47.3
B	28.9	26.6	39.0	27.9
C	35.5	16.4	43.3	27.2
D	38.7	37.7	51.8	37.3

Results for the second focus of our research are in Table 3. The ASR output of the topic-labelled subset of the Finnish lecture speech data was used to automatically classify the dominant topic during two-minute segments in the recordings. In total, there were nine pre-defined topics, with specific keywords attached to each topic. The evaluation method utilised was quite rudimentary in nature, but the results do indicate that, even with keyword error rates of over 30%, this topic-classification method still has the potential to be accurate.

Table 3. Results for the topic-labelled subset of the Finnish teacher speech data (5 h 18 min). The results are reported via word error rate (WER), keyword error rate (KER) and topic-classification accuracy.

ASR system	WER [%]	KER [%]	Topic classification [%]
Kaldi toolkit	33.0	32.3	84.5
Google Speech API	47.5	35.8	85.7

DISCUSSION AND CONCLUSION

Our first research question was: How well does ASR perform when interpreting lesson speech? In particular, what is the keyword error rate? The results for the Finnish recordings have been given in Tables 1 and 2 above. As discussed, we achieved better ASR accuracy with the Kaldi toolkit than with the Google Speech API. In order to achieve even better recognition accuracy, we would need to include more language-model training data. If our aim, however, is to assist researchers with writing the transcriptions, then the error rate of 36.3% could be considered acceptable. The result does not mean that automatically produced transcripts are missing 36.3% of the words; although some words are missing, words containing one wrong or missing letter are also counted as errors, thus making the error percentage appear higher.

Our second research question was: To what extent can ASR can be used for the automatic recognition of lesson topics? The results are presented in Table 3 above. We found that, although the error rates given for single words were over 30%, the ASR system still could recognise lesson topics most of the time. It can be said, therefore, that topic recognition is not particularly susceptible to single-word errors. These results are promising and they indicate the potential of ASR in the study of classroom talk.

Our research project is still in its early stages. More data needs to be collected from the classroom environment in order to improve the ASR system. In addition, the mobile application needs to be tested more extensively in the classroom in order to develop its functionality.

Going forward, the aim is to develop a mobile application that supplies feedback in the form of charts and graphs to teachers, quickly and easily after the observed lesson thus allowing them to review their own teaching. It is probable that the ASR aspect will still have to be conducted via computer in future experiments, but the application will be available for use in recording speech and making classroom observations. This would be a great tool for in-service teachers' education because it would enable teachers to recognise their habits and make changes where appropriate.

Our long-term aim is to develop an ASR system with the capability of analysing areas such as teacher questioning automatically. The transcriptions of teacher talk could facilitate the classification of the teachers' discourse (authoritative or interactive etc.). The transcripts could be compared with observations to provide further insights. It is hoped that, once a large database of lesson transcripts and manual classifications has been gathered, machine-learning algorithms could be used to suggest classifications automatically. Elsewhere, we have developed automatic visualisations of conceptual networks in physics lessons and achieved promising results (Caballero et al., 2017). Such networks could help teachers to make connections between different keywords, thus facilitating interconnectivity in the teaching of concepts. The data would also enable teachers to see how widely teachers they use various concepts during their classes. In this way, classifications of teacher discourse, alongside the automatic recognition of questions and keywords, would provide immediate support to researchers and teachers.

REFERENCES

Aller Media ltd. (2014). *The Suomi 24 Corpus (2016H2)* [text corpus]. Kielipankki. Retrieved from http://urn.fi/urn:nbn:fi:lb-2017021506

Abrahams, I. & Millar, R. (2008). Does Practical Work Really Work? A study of the effectiveness of practical work as a teaching and learning method in school science. *Journal of Science Education*, 30(14), 1945–1969.

Bahl, L. & Jelinek, F. (1975). Decoding for channels with insertions, deletions and substitutions with application to speech recognition. *IEEE Transactions on Information Theory*, 21, 404–411.

Blanchard, N., Brady, M., Olney, A. M., Glaus, M., Sun, X., Nystrand, M., Samei, B., Kelly, S., & D'Mello, S. (2015). A Study of Automatic Speech Recognition in Noisy Classroom Environments for Automated Dialog Analysis. *Artificial Intelligence in Education*, 9112, 23-33.

Blei, D., Jordan, Ng, A. Y., & Jordan, M. (2003). "Latent Dirichlet Allocation". *Journal of Machine Learning Research,* **3** (4–5), 993–1022.

Caballero, D., Araya, R., Kronholm, H., Viiri, J., Mansikkaniemi, A., Lehesvuori, S., Virtanen, T. & Kurimo, M. (2017). ASR in Classroom Today: Automatic Visualization of Conceptual Network in Science Classrooms. In É. Lavoué et al. (Eds.): *EC-TEL 2017, LNCS 10474*, pp. 541–544, 2017. Springer International Publishing AG 2017. DOI: 10.1007/978-3-319-66610-5_58.

Deerwater, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R., (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6), 391- 407.

Driver, R., Asoko, H., Leach, J., Scott, P., & Mortimer, E. (1994). Constructing scientific knowledge in the classroom. *Educational Researcher*, 23(7), 5-12.

Google. Speech API - speech recognition. Retrieved from https://cloud.google.com/speech/

Enarvi, S., & Kurimo, M. (2013). *Studies on Training Text Selection for Conversational Finnish Language Modeling*. In Proceedings of the 10th International Workshop on Spoken Language Translation.

Helaakoski, J. & Viiri, J. (2014). Content and content structure of physics lessons and students' learning gains. In H.E. Fischer,P. Labudde, K. Neumann, & J. Viiri, (2014) *Quality of instruction in physics. Comparing Finland, Germany and Switzerland.* Waxmann, Munster. pp. 93-110.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A.,Vanhoucke, V., Nguyen, P., Sainath, T., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine*, 29, 82–97.

Iskra, D. J., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., & Kiessling, A. (2002). Speecon-speech databases for consumer devices: Database specification and validation. In LREC, 2002.

Juang, B. H., Levinson, S., & Sondhi, M. (1986). Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE Transactions on Information Theory*, 32, 307–309.

Kiemer, K., Gröschner, A., Pehmer, A., & Seidel, T. (2015). Effects of a classroom discourse intervention on teachers' practice and students' motivation to learn mathematics and science. *Learning and Instruction, 35*, 94-103.

Leach, J., & Scott, P. (2003). Individual and sociocultural views of learning in science education. *Science and Education*, 12(1), 91-113.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G. & Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In Interspeech 2016, pp. 2751–2755. [Online]. http://dx.doi.org/10.21437/Interspeech.2016-595

Ranchal, R., Taber-Doughty, T., Guo, Y., Bain, K., Martin, H., Robinson, J. P., & Duerstock, B. S. (2013). Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom. *IEEE Transactions on Learning Technologies*, 6(4), 299-311.

Sapru, A., & Bourlard, H. (2014). Detecting speaker roles and topic changes in multiparty conversations using latent topic models. Proceedings of Interspeech 2014.

Spärck Jones, K., A. (1972). Statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28 (1), 11-21.

Wang, Z., Pan, X., Miller, K. F., & Cortina, K. S. (2014). Automatic classification of activities in classroom discourse. *Computers & Education*, 78, 115-123.

Yamamoto, N., Ogata, J., & Ariki, Y. (2003). Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition. Proceedings of Interspeech 2003.

Young, S., Russell, N., & Thornton, J. (1989). Token passing: A simple conceptual model for connected speech recognition systems. Technical Report CUED/F-INFENG/TR.38, Cambridge University Engineering Department.