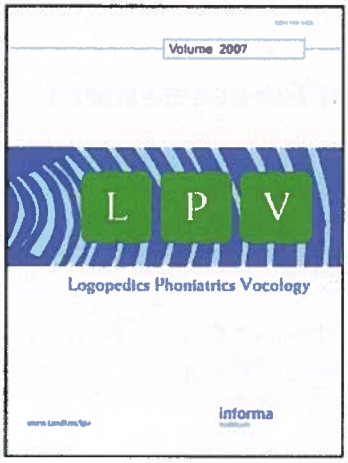
Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice. speech, and language disorders

Kitzing, Peter; Maier, Andreas; Lyberg Åhlander, Viveka

Published in:
Logopedics Phoniatrics Vocology

DOI:
10.1080/14015430802657216

2009



Logopedics Phoniatrics Vocology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713713058

Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders

Peter Kitzing [a]; Andreas Maier [b]; Viveka Lyberg Åhlander [c]
[a] Faculty of Engineering, Department of Design Sciences, Institute of Rehabilitation Engineering Research (Certec), Lund University, Sweden [b] Chair of Computer Science 5 (Pattern Recognition) of the Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany [c] Faculty of Medicine, Logopedics, Phoniatrics, Audiology, Lund University, Sweden

First Published on: 28 January 2009

To cite this Article Kitzing, Peter, Maier, Andreas and Åhlander, Viveka Lyberg(2009)'Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders',Logopedics Phoniatrics Vocology,
To link to this Article: DOI: 10.1080/14015430802657216
URL: http://dx.doi.org/10.1080/14015430802657216

Logopedics Phoniatrics Vocology
2009, 1–6, iFirst article

informa
healthcare

LPV FORUM

Automatic speech recognition (ASR) and its use as a tool for assessment or therapy of voice, speech, and language disorders

PETER KITZING[1], ANDREAS MAIER[2] & VIVEKA LYBERG ÅHLANDER[3]

[1]*Faculty of Engineering, Department of Design Sciences, Institute of Rehabilitation Engineering Research (Certec), Lund University, Sweden,* [2]*Chair of Computer Science 5 (Pattern Recognition) of the Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany, and* [3]*Faculty of Medicine, Logopedics, Phoniatrics, Audiology, Lund University, Sweden*

Abstract
In general opinion computerized automatic speech recognition (ASR) seems to be regarded as a method only to accomplish transcriptions from spoken language to written text and as such quite insecure and rather cumbersome. However, due to great advances in computer technology and informatics methodology ASR has nowadays become quite dependable and easier to handle, and the number of applications has increased considerably. After some introductory background information on ASR a number of applications of great interest for professionals in voice, speech, and language therapy are pointed out. In the foreseeable future, the keyboard and mouse will by means of ASR technology be replaced in many functions by a microphone as the human–computer interface, and the computer will talk back via its loud-speaker. It seems important that professionals engaged in the care of oral communication disorders take part in this development so their clients may get the optimal benefit from this new technology.

Key words: *Automatic speech recognition, ASR, computer-aided language learning, CALL, computer-aided pronunciation training, CAPT, computer-aided speech therapy, voice quality assessment*

In the early days of information technology (IT), more than 25 years ago, the senior author was engaged in developing software for computer-aided aphasia therapy. As a rule, the patients were enthusiastic about the computer being able not only to present language exercises but also to provide feedback as to the quality of the users' achievements, be it in speech comprehension or reading and writing. Only when it came to the feedback on spoken utterances, the computer failed. The user had to depend on his own self-monitoring or an evaluation by his speech/language therapist or some other person. Rather often, the aphasic users complained about the lack of computer control for speech exercises. Problems when speaking was their greatest loss, and adequate speech exercises were what they wanted most.

Why was it that the computer could accept and adequately correct written material but not speech? The reason is that, contrary to printed or typed text, oral speech—even if basically correct as to its pronunciation—varies widely from person to person and also for the same speaker on different occasions. Of course, a computer could easily have recognized exact copies of some acoustic recordings. But already a small difference of a copy from the original, e.g. a change in pitch, would cause the computer to reject it entirely, even if it were perfectly understandable to a human listener. One way to overcome this obstacle would have been to use a dictation system as input to the computer. However, the early software for speech recognition was too cumbersome to handle, too slow, and too undependable.

Automatic speech recognition (ASR) systems started to appear in the middle of the last century, but they were not very successful in the beginning. Their performance was limited to the recognition of single words like isolated digits, which had to be pronounced with a distinct pause between them (1). In 1984 IBM had developed a system that could

Correspondence: Peter Kitzing, MD, PhD, Certec, LTH, Box 118, S-22100 Lund, Sweden. E-mail: peter.kitzing@med.lu.se

(Received 5 June 2008; revised 6 October 2008; accepted 29 October 2008)

ISSN 1401-5439 print/ISSN 1651-2022 online
DOI: 10.1080/14015430802657216

recognize a vocabulary of about 5000 English words (when spoken with small pauses in between), but it needed several minutes of computing time on a large desktop computer. The aim of many commercial applications of ASR was to replace the job of medical transcriptionists, but the systems were generally not accepted, mostly because of shortcomings of the software. The continuous speech of narrative dictation proved to be too difficult to achieve sufficiently accurate results, and the large amount of time required by the user to train the software did not seem justified.

ASR systems may be classified according to their capacity to handle different factors that influence the signal to be decoded, see Table I (2). Classification of ASR systems often begins by dividing them into speaker-dependent versus speaker-independent, the former usually combined with a large vocabulary and the latter with a smaller one. In speaker-dependent systems each user has to train the system by reading aloud for a certain time to create an individual so-called profile. Because of differences in sex and age, social background, and personal physical or emotional state, etc., speakers vary as to their voice quality and articulation but also in their general speaking behaviour. One and the same linguistically coded utterance (=written sentence) may therefore be phonetically realized in a huge number of different ways for the system to decipher correctly. Last but not least, the analysis may be hampered by degradation of the acoustic signal caused by ambient noise, insufficient microphone quality, or even faulty microphone-to-mouth distance.

In spite of many large obstacles the development of ASR systems has been a great success so that the systems nowadays have become more speaker-independent and they accept larger vocabularies. There has been a great increase in the number of applications, some of which will be discussed below. Several factors have played a significant part in bringing about such progress (3). One important factor is the introduction during the last decades of more efficient statistical algorithms based, nowadays almost exclusively, on the use of the *hidden Markov models* (HMM), allowing the systems to be trained automatically. One prerequisite for such training (and also testing) is the development of *large speech corpora* having taken place, not infrequently comprising tens of thousands of sentences. Another factor is the establishment of *standards* for speech corpus and system performance *evaluation*, so that test results published from different research groups become comparable. Last but not least, the fast development of *computer technology* with its continuously increasing storage capacities and processing speeds has been of great importance. The amount of computation bought for a certain cost has been doubled in every 12–18 months (Moore's law).

Table I. Factors influencing the function of ASR systems (modified from Zue et al. 1996 (2)).

Enrolment	speaker-dependent versus speaker-independent
Vocabulary	small (<20 words) versus large (>20,000 words)
Speaking mode	isolated words versus continuous speech
Speaking style	reading aloud versus spontaneous speech
Language model	finite state versus context-sensitive
Perplexity (~ word variation in a text)	small (radiology) versus large (journalism, general English)
Signal to noise ratio (SNR), (e. g. ambient noise)	high (>30 dB) versus low (<10 dB)
Transducer	microphone quality, microphone placement

The function of a typical ASR system proceeds in different steps. The acoustic speech signal, consisting of variations of air pressure, is transformed into a varying electric current by a microphone. After adjustment to constant volume, the electric signal is filtered to avoid unwanted noise and aliasing. It is then analogue-to-digital converted (typical sampling frequencies 8–25 kHz, 16 bit) and framed or 'cut into slices' by so-called (overlapping) Hanning or Hamming windows (frame rate 50–100 Hz, window length 16–32 ms). The resulting frames or signal 'slices' can be further processed as (almost) stationary entities. They are then acoustically analysed, and according to auditory perceptual function they may be coded with Mel frequency cepstral coefficients (MFCC) (4) or perceptual linear prediction (PLP) coefficients (5), their first and second order derivatives, as well as normalized by vocal tract length normalization (VTLN) (6). By feature extraction the frames are then categorized according to their spatial qualities and coded to reduce the amount of data and to allow the subsequent pattern matching and recognition process. Data sampled at 16 kHz, for example, results in a window size of 512 samples from which about 13 MFCCs are extracted. With first and second order derivatives this yields a feature vector of dimension 39. Hence the reduction is of an order of magnitude. The resulting mathematical speech frame information is compared with similar speech frames from a training corpus of speech. By means of mathematical-statistical methods so-called Gaussian mixture models (GMMs) are included into the hidden Markov models (HMM). The measures of similarity (or matching distance) can then be calculated as the probability of a certain feature

vector given a set of HMMs (7). Best matches are chosen as candidates to form speech sounds or phones (=phonetic realizations of phonemes, including variants, e.g. due to coarticulation). By analogical statistical methods phones are put together (concatenated) to form words, and words may be concatenated to create sentences. These processing steps depend much on adequate language models, both at the phoneme, word, and sentence level. Language models may be either statistically or linguistically grounded.

A mathematically thorough description of the process is beyond the scope of this paper. On the other hand, maybe a simple metaphor may give some notion of the method and be useful as an answer to the initial question why computers need complicated ASR systems to recognize speech even if they easily can tell right from wrong in written material. In a way, this is a similar difference as that between a memory game and a jig-saw puzzle. In memory games a picture can only be entirely right or wrong. But in a puzzle the picture is built up by many pieces, some of which sometimes may be 'almost right'. If such a piece is put into place with force, it can be seen not to compare well with the rest. By analogy, ASR systems do not always produce one hundred per cent correct results.

The quality of ASR systems may be expressed by the word error rate (WER or just E). The sum of substitutions (S), insertions (I), and deletions (D) divided by the total number of words to be recognized is multiplied by 100, resulting in the percentage of WER.

$$E = 100 \times (S + I + D)/N \quad (1)$$

An equivalent measure of correct function is word accuracy (WA),

$$WA = 100 \times (NC - NW)/N \; (= 100 - E) \quad (2)$$

where N is the total number of words to be processed, NC the number of correctly recognized words, and NW the number of wrongly inserted words (=I above). Other quality criteria are recognition speed, vocabulary size, and degree of speaker dependence.

Besides commercially available ASR products for transcription of speech to text, the use of ASR in conversational telephone dialogue systems may be mentioned. These systems handle speaker-independent speech recognition via the telephone line. Therefore, they have to cope with effects such as varying signal quality—depending on the transmission channel—and reduced quality of telephone speech (4 kHz bandwidth). Modern systems handle small to medium-sized vocabularies with up to 10,000 words reliably and can even handle words that did not appear in the vocabulary (8).

The digital speech to text function is useful not only for medical transcription but also in court reporting and television subtitling (9), in visual voice mail services for stationary and portable phones (e.g. for deaf people to read what they cannot hear), and as a basis for computerized automatic translation. Others who may benefit from the possibility of oral input to the computer are people with dyslexia or those with difficulties using their hands, be it from serious medical deficiencies or only slight repetitive stress injuries. ASR systems may also be used for voice command recognition, e.g. in industrial or military settings like fighter aircraft or in battle management command centres.

Not only can modern ASR systems produce text from speech, but they are also able to supply a scoring of their performance. This is because of the recognition process being a sort of mathematical comparison of probabilities resulting in a measure of likelihood or distance between the actual performance and a model. This facility is used to construct evaluation and feedback functions in programs for computer-assisted language learning (CALL) (10) and in computer-aided assessment and therapy for voice, speech, and language disorders (11).

In the field of second language (L2) learning the use of ASR has mainly focused on computer-aided pronunciation training (CAPT). Earlier software for such a function more or less mimicked the classic language labs with recorded examples for the students to imitate and with the comments of a human listener (e.g. the teacher) as only feedback. Self-monitoring would not be enough, because most L2 learners listen 'phonemically' according to their own L1 system and not to their actual phonetic realization of the speech sounds, so own attention is not enough for them to become aware of their mispronunciations even if they can listen to a model. Feedback as in earlier CAPT systems by visual displays of phonetic registrations like waveforms, spectrograms, or intonation curves was not very helpful, as the most serious pronunciation errors occurred on the segmental level not easy to spot on such displays. To be of real help, the CAPT system should be able to produce online speech quality scores and to offer examples and training exercises for corrections. Many obstacles have to be overcome to achieve this goal, and the problem is multifaceted (12). The first focus should not be on the technical facilities but on the needs of the user. Does (s)he want to pass a test of general speech proficiency as it is administered to immigrants to the Netherlands within the framework of the Newcomer Integration Act (13)? Does (s)he want to learn just intelligibility

not bothering with a slight accent, or is the goal an entirely correct pronunciation (as e.g. needed by a spy)? Should the system focus on global or segmental scoring? Which people are relevant to be recruited as evaluation experts and on which criteria do they found their assessments? How do their scorings vary from rater to rater and from one time to another by the same rater? And lastly, and after all maybe most importantly: How should the ASR software be designed?

Research on these topics has grown to an extent too large to be reported here. May it suffice to point to just one project as an illustration of the potential of modern technology in the field. The project is called the Virtual Language Tutor (14), and it is intended for remediation of articulation deficiencies. Its ASR system can recognize utterances even if they are not correctly pronounced, and by means of an additional software module, the Pronunciation Analyzer, it can show mispronounced phonemes on the visual display to the speaker. The system also includes an embodied communication agent (ECA) in the form of a 3D so-called talking head, controlled by the software for text-to-visual-speech synthesis. Visual feedback of the movements of the articulators is also provided.

The use of sophisticated systems for ASR-assisted speech training is of course not restricted only to second language (L2) learning but may also be applied in the treatment of speech disorders. One such system is the Ortho-Logo-Paedia (OLP) computer program, which includes motivating games as a feedback for children to target difficult speech sounds. Targets may be individualized based on the user's best performance rather than on the standard pronunciation of the language, which may be too hard to accomplish in severe dysarthria. It has been shown, however, that even in such cases, OLP-guided exercises may be useful. Seven adult subjects with long-standing dysarthria of different aetiologies carried out OLP-guided exercises during two blocks of 6 weeks. All showed improvement, and it was shown that this was as effective as traditional therapy (15). The OLP has been successfully used also in speech and language therapy for children with hearing impairment (16). An interesting use of ASR systems to aid very degraded articulation is reported by Parker (17). Even if practically unintelligible, severely dysarthric speech may contain some characteristic information carrying elements which by means of special ASR systems could be identified and used by the subjects to operate their environmental control systems (ECSs) like e.g. the remote control of their TV set. Before, subjects were taught to use their remaining identifiable strings of articulation more consistently.

Whether ASR systems are to be used more generally in future didactic or remedial pedagogics depends to some degree on how their distribution and integration in the already existing activities of language learning and speech therapy are organized. The Computer-Assisted Language Learning from Erlangen (CALLER) project is an approach to solve this important question (18). Texts and exercises as well as a protocol (for the teacher) of the students' work are stored on a central server, whereas the actual program, including animations and an ECA, is run on local lap-tops in the schoolroom or at home.

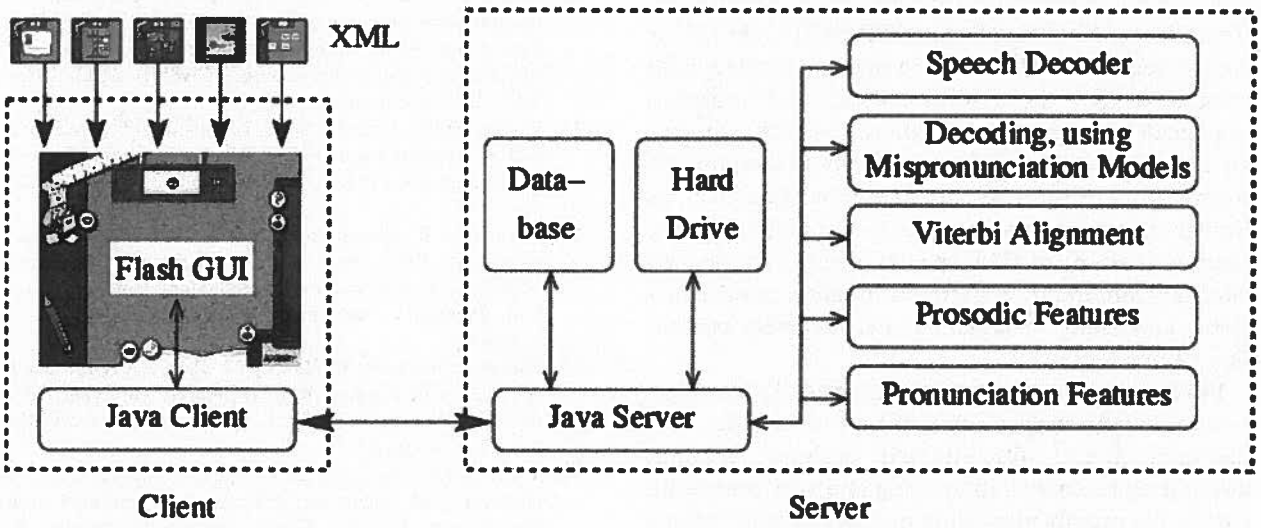


Figure 1. Client-server architecture of the CALLER system (18). New exercises and vocabulary can be added to the system dynamically via the XML interface. While the exercises and the graphical user interface run on the client's computer, all performance-intensive computations are carried out on the server.

A similar structure for administration has been used in the Program for Evaluation and Analysis of all Kinds of Speech Disorders (PEAKS) (19). Speech recordings to be assessed can be transferred via a secure connection on the Internet or by telephone to a central server including an ASR system for analyses. The results are sent back to the client, and the recordings are stored in a database. At the present time, data are collected from patients with insufficient palatal function (CLP) (11), tracheoesophageal (shunt) speech (TE) (20), and early oral or laryngeal cancer (19). The reason for this selection of diagnoses is obviously that they cause very significant alterations of speech quality and that in earlier studies ASR-assisted quality scorings of such speech correlated very well with evaluation by expert listeners. Among other noteworthy results of this project and its preceding studies may be mentioned the high consistency among expert listeners when evaluating defective speech and the possibility of easy Internet or telephone data transfer (21). Achievements as to the ASR methodology are the surprisingly good results even with speech degraded by telephone transfer; that inclusion of paralinguistic and suprasegmental information in the recognition process yields more robust results; and that instead of transliteration ordinary text reference is sufficient for probability evaluation (20).

It should be pointed out, however, that all the here-mentioned results of speech quality only refer to a global, continuous aspect of speech. They are not applicable to linguistic segments like words (aphasia), and they are not differentiated as to different physiological, acoustic, or auditory criteria of speech and especially voice quality. For example, only the grade (G) aspect in the Grade, Roughness, Breathiness, Asthenia, Strain (GRBAS) scale (22) is considered. G is well known to mainly correlate with roughness (R), the typical indicator of morpho-organic deficiencies to be diagnosed as such, whereas an essentially functional component like strain (S) seems to contribute less to the global G aspect. A similar discussion may pertain to the global scoring (and evaluation) of CLP speech, where experienced clinicians differentiate between nasality, nasal emissions, and faulty articulation, i.e. incorrect realization of phonemes.

Further development of ASR scoring of pathologic voice function and speech will have to aim at more fine-grained and differentiated analyses. Without doubt it represents an interesting and very promising part of the rapidly increasing number of applications of automatic speech recognition.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

1. Jelinek F. Continuous speech recognition by statistical methods. IEEE Proceedings. 1976;64:532–56.
2. Zue V, Cole R, Ward W. Speech recognition. In: Cole RA, editor. Survey of the state of the art in human language technology. Cambridge: Cambridge University Press; 1996.
3. Furui S. 50 years of progress in speech and speaker recognition. In: Proceedings of the 10th International Conference on Speech and Computer (SPECOM). Patras, Greece: Conference Proceedings by University of Patras; 2005. p. 1–9.
4. Davis SB, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust. 1980;28:357–66.
5. Furui S. Digital speech processing. 2nd ed. New York, USA: Marcel Dekker; 2000.
6. Stemmer G. Modeling variability in speech recognition. Berlin, Germany: Logos Verlag; 2005.
7. Huang X, Huang A, Acero A, Hon H-W. Spoken language processing—a guide to theory, algorithm, and system development. Upper Saddle River: Prentice Hall; 2001.
8. Gallwitz F. Integrated stochastic models for spontaneous speech recognition. In: Studien zur Mustererkennung, vol. 6. Berlin, Germany: Logos Verlag; 2002.
9. Lambourne A, Hewitt J, Lyon C, Warren S. Speech-based real time subtitling service. International Journal of Speech Technology. 2004;7:269–79.
10. Hacker C, Batliner A, Steidl S, Nöth E, Niemann H, Cincarek T. Assessment of non-native children's pronunciation: human marking and automatic scoring. In: Proceedings of the 10th International Conference on Speech and Computer (SPECOM). Patras, Greece: Conference Proceedings by University of Patras; 2005. p. 61–4.
11. Schuster M, Maier A, Haderlein T, Nkenke E, Wohlleben U, Rosanowski F, et al. Evaluation of speech intelligibility for children with cleft lip and palate by automatic speech recognition. Int J Pediatr Otorhinolaryngol. 2006;70:1741–7.
12. Neri A Cuchiarini C Strik C Feedback in computer assisted pronunciation training: technology push or demand pull? In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2002. Orlando, USA: IEEE Computer Society Press; 2002. p. 1209–12.
13. Cucchiarini C, Strik H, Boves L. Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. J Acoust Soc Am. 2002;111:2863–73.
14. Granström B. Speech technology for language training and e-inclusion. Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech). Lisbon, Portugal: Conference Proceedings by ISCA; 2005. p. 449–52.
15. Palmer R, Enderby P, Hawley M. Addressing the needs of speakers with longstanding dysarthria: computerized and traditional therapy compared. Int J Lang Commun Disord. 2007;1 42 Suppl:61–79.
16. Öster A-M. Computer based speech therapy using visual feedback with focus on children with profound hearing impairments [Doctoral Thesis]. Stockholm, Sweden: KTH Computer Science and Communication; 2006.

17. Parker M, Cunningham S, Enderby P, Hawley M, Green P. Automatic speech recognition and training for severely dysarthric users of assistive technology: the STARDUST project. Clin Linguist Phon. 2006;20:149–56.
18. Hacker C, Maier A, Hessler A, Guthunz U, Nöth E. Caller: computer assisted language learning from Erlangen—pronunciation training and more. International Conference on Interactive Computer Aided Learning (ICL), Villach, Austria. Kassel, Germany: Kassel University Press; 2007.
19. Maier A, Schuster M, Batliner A, Nöth E, Nkenke E. Automatic scoring of the intelligibility in patients with cancer of the oral cavity. In: Interspeech 2007— Antwerp, Belgium: Conference Proceedings by ISCA, 27–31 August 2007, Antwerp, Belgium. p. 1206–9.
20. Haderlein T. Automatic evaluation of tracheoesophageal substitute voices. Berlin, Germany: Logos Verlag; 2007.
21. Riedhammer K, Stemmer G, Haderlein T, Schuster M, Rosanowski F, Nöth E, et al. Towards robust automatic evaluation of pathologic telephone speech. Proceedings of the Automatic Speech Recognition and Understanding Workshop (ASRU). Kyoto, Japan: IEEE Computer Society Press; 2007. p. 717–22.
22. Hirano M. Clinical examination of voice. New York: Springer, 1981.