Susanne Wagner (Halle)

Intralingual speech-to-text-conversion in real-time: Challenges and Opportunities

Contents

Abstract

Intralingual speech-to-text-conversion is a useful tool for integrating people with hearing impairments in oral communication settings, e. g. councelling interviews or conferences. However, the transfer of speech into written language in real time requires special techniques as it must be very fast and almost 100% correct to be understandable. The paper introduces and discusses different techniques for intralingual speech-to-text-conversion.

1 The need for real-time speech-to-text conversion

Language is a very fast and effective way of communicating. To use language means to express an unlimited amount of ideas, thoughts and practical information by combining a limited amount of words with the help of a limited amount of grammatical rules. The result of language production processes are series of words and structure. Series of words are produced – i.e. spoken or signed – in a very rapid and effective way. Any person can follow such language production processes and understand what the person wants to express if two preconditions are fulfilled the recipients must:

1. know the words and grammatical rules the speaker uses and
2. be able to receive and process the physical signal.

Most people use oral language for everyday communication, i.e. they speak to other people and hear what other people say. People who are deaf or hard-of-hearing do not have equal access to spoken language, for them, precondition 2 is not fulfilled, their ability to receive speech is impaired.

If people who are severely impaired in their hearing abilities want to take part in oral communication, they need a way to compensate their physical impairment[1]. Hearing aids are sufficient for many hearing impairment people. However, if hearing aids are insufficient,

[1] To provide access to oral communication situations for hearing impaired people is an issue of fairness which, in recent years, is increasingly reflected by national governments. In some countries laws stipulate that at least authorities and official institutions provide information in a form which is also accessible for people with an impairment. Consequently, auditory information has to be provided in a way which can also be detected visually or haptically by people with a hearing impairment (cf. S. Wagner et al., 2004).

spoken language has to be transferred into a modality which is accessible without hearing, e.g. into the visual domain.

There are two main methods to transfer auditory information into a visible format. The translation into sign language is one method and it is best for people who use sign language as a preferred language, as e.g. many Deaf people do. However, for people with a hearing disability who do not know sign language, sign language interpreting is not an option — as for many Hard of Hearing people and people who became hearing impaired later in their life or elderly people with various degrees of hearing loss. They prefer their native oral language given in a visible modality. For them, a transfer of spoken words into written text is the method of choice, in other words: they need an intralingual speech-to-text-conversion.

Speech-to-text-translation (audiovisual translation) of spoken language into written text is an upcoming field since movies on DVDs are usually sold with subtitles in various languages. While the original language is given auditorily, subtitles provide a translated version in another language at the same time visually. The audiovisual transfer from the spoken original language into other languages which are presented in the subtitles can be called an interlingual audiovisual translation. Interlingual translation aims at transferring messages from one language into another language. This translation process combines classical interpreting with a transfer from spoken language patterns into written text patterns. Auditory events which are realized as noises or speech melodies would often not be transferred because normally hearing people can interpret them by themselves. Interlingual translation primarily addresses the lack of knowledge of the original language, i.e. the first precondition for understanding language.

The intralingual audiovisual transfer differs in many aspects from the interlingual audiovisual translation between two languages.

First of all, intralingual audiovisual transfer for people with hearing impairments addresses primarily precondition 2, i.e. the physical ability to perceive the speech signals. The aim of an intralingual audiovisual transfer is to provide all auditory information which is important for the understanding of an event or action. Words as well as non-language sounds like noises or hidden messages which are part of the intonation of the spoken words (e.g. irony or sarcasm) need to be transmitted into the visual (or haptic) channel. How this can be achieved best, is a question of present and future research and development (cf. Neves, in this book). Moreover, people with hearing impairment may insist on a word-by-word-transfer of spoken into written language because they do not want a third person to decide which parts of a message are important (and will therefore be transferred) and which parts are not. As a result, intralingual audiovisual transfer for people with hearing impairment might mean that every spoken word of a speech has to be written down and that all relevant auditory events from outside of the speech have to be described, too (interruptions, noises). In the latter case, the intralingual audiovisual transfer would exclusively satisfy the physical ability to perceive the speech signal (precondition 2).

The classical way to realize an intralingual speech-to-text transfer is to stenotype a protocol or to record the event and to transfer it into a readable text subsequently. This post-event transfer process is time-consuming and often difficult, since auditory events easily become ambiguous outside of the actual context. Moreover, the time shift involved in the transfer into a readable text means a delayed access to the spoken words, i.e. it does not help people with hearing impairments in the actual communication situation. However, for counselling interviews, at the doctor's or at conferences, access to spoken information must be given in real-time. For these purposes, the classical methods do not work.

2 The challenges of speech-to-text-conversion in real-time

Real-time speech-to-text-conversion aims at transferring spoken language into written text (almost) simultaneously. This gives people with a hearing impairment, access to the contents of spoken language in a way that they e.g. become able to take part in a conversation within the normal time frame of conversational turn taking. Another scenario for real-time speech-to-text-transfer is a live broadcast of a football match where the spoken comments of the reporter are so rapidly transferred into subtitles that they still correspond to the scene the reporter comments on. An example from the hearing world would be a parliamentary debate which ends with the electronic delivery of the exact word protocol presented to the journalists immediately after the end of the debate. (cf. Eugeni, forthcoming)

This list could be easily continued. However, most people with a hearing disability do not receive real-time speech-to-text services at counselling interviews, conferences or when watching a sports event live on TV. Most parliamentary protocols are tape recorded or written stenotyped and subsequently transferred into readable text. What are the challenges of real-time speech-to-text conversion that make its use so rare?

2.1 Time

A good secretary can type about 300 key strokes (letters) per minute. Since the average speaking rate is about 150 words per minute (with some variance between the speakers and the languages), even the professional typing rate is certainly not high enough to transfer a stream of spoken words into a readable form in real-time. As a consequence, the speed of typing has to be increased for a sufficient real-time speech-to-text transfer. Three different techniques will be discussed in the following section “methods”.

2.2 Message Transfer

The main aim of speech-to-text transfer is to give people access to spoken words and auditory events almost simultaneously with the realization of the original sound event. However, for people with limited access to spoken language at a young age, 1:1 transfer of spoken words into written text may sometimes not be very helpful. If children are not sufficiently exposed to spoken language, their oral language system may develop more slowly and less effectively compared with their peers. As a result, many people with an early hearing impairment are less used to the grammatical rules applied in oral language as adults and have a less elaborated mental lexicon compared with normal hearing people (Schlenker-Schulte, 1991; see also Perfetti et al. 2000 with respect to reading skills among deaf readers)[2].

If words are unknown or if sentences are too complex, the written form does not help their understanding. The consequence for intralingual speech-to-text conversion is that precondition 1, the language proficiency of the audience, also has to be addressed, i.e. the written transcript has to be adapted to the language abilities of the audience - while the speech goes on.

Speech-to-text service providers not only need to know their audience, they also have to know which words and phrases can be exchanged by equivalents which are easier to

[2] Apart from people who were born with a more severe hearing impairment, language proficiency might differ also for people with cultural backgrounds different from a majority group, people with other mother tongues or people with learning difficulties.

understand, and how grammatical complexity can be reduced. They need to know techniques of how to make the language in itself more accessible while the information transferred is preserved. Aspects of how language can be made more accessible will be discussed in the following section "text adaptation".

2.3 Real-time presentation of the written text

Reading usually means that words are already written down. Presented with a written text, people will read at their individual reading speed. This, however, is not possible in real-time speech-to-text conversion. Here, the text is written and read almost simultaneously, and the control of the reading speed shifts at least partly over to the speaker and the speech-to-text provider. The text is not fixed in advance, instead new words are produced continuously and readers must follow this word production process very closely if they wants to use the real-time abilities of speech-to-text transfer. Because of this interaction of writing and reading, the presentation of the written text must be optimally adapted to the reading needs of the audience. This issue will be discussed at the end of the paper in section "presentation format".

The challenges of real-time speech-to-text conversion can now be summarized as follows:

1. to be fast enough in producing written language that
2. it becomes possible to meet the expectations of the audience with respect to the characteristics of a written text. Word-by-word transfer enhanced by a description of auditory events from the surroundings as well as adaptations of the original wording into easier forms of language must be possible. Moreover,
3. a successful real-time presentation must match the reading abilities of the audience, i.e. the written words must be presented in a way that is optimally recognizable and understandable for the readers.

3 Methods of real-time speech-to text conversion

There are three methods that are feasible when realizing (almost) real-time speech-to-text transfer: speech recognition, computer assisted note taking (CAN) and communication access (or computer aided) real-time translation (CART). The methods differ

1. in their ability to generate exact real-time transcripts.
2. with respect to the conditions under which these methods can be properly applied and
3. with respect to the amount of training which is needed to become a good speech-to-text service provider.

3.1 Speech recognition

Automatic speech recognition (ASR) technologies today can correctly recognize and write down more than 90% percent of a long series of spoken words for many languages. However, even this high percentage is not sufficient for speech-to-text services, since 96+x% correctness is needed to provide a sufficient message transfer (Stinson et al. 1999: accuracy). Moreover, even the 90+x% accuracy in automatic speech recognition does not occur by itself. In order to be recognized, the speaker has to train the speech recognition system in advance with her/is voice and speaking characteristics. Some regional speaking characteristics (dialects) are generally only poorly recognized, even after extensive training. Physical changes in voice quality (e.g. from a flu) can result in poorer recognition results. The reason for this is that the speech recognition process is based on a match of physical parameters of the actual speech signal with a representation which was generated on the basis of a general

phonetic model of language and the phonetic and voice data from the individual training sessions. If the individual physical parameters differ from those of the training sessions, recognition is less successful. Moreover, if background noise decreases the signal-to-noise-ratio, accuracy might go down to below 80 percent.

However, speech recognition systems can meet challenge number 1 (writing speed) under good circumstances. In this case, the recognition rate of ASR would in principle be high enough to transfer every spoken word into written text in real-time. But there are limitations which have to be taken into account. The most restrictive factor is that automatic speech recognition systems are not (yet) capable of recognizing phrase- and sentence boundaries (but see Leitch et al. 2002). Therefore, the output from an automatic speech recognition system is a stream of words without any comma or full stop. Moreover, the words would not be assigned to the different speakers. An example from Stuckless (1999) might illustrate how difficult it is to understand such a stream of words:

> "why do you think we might look at the history of the family history tends to dictate the future okay so there is some connection you're saying what else evolution evolution you're on the right track which changes faster technology or social systems technology." (Stuckless 1999)

Automatic speech recognition today fails as far as challenge 3 is concerned.: Although the single words are readable, the output of automatic speech recognition systems is almost not understandable for any reader.

The short-term solution for this problem is that a person, who has trained her/is speech recognition system extensively with his/her speaking characteristics, has to re-speak the speech of the speaker with explicit punctuation commands and speaker identification. With re-speaking, speech recognition is an option especially for live subtitling and conferences where the speech-to-text conversion can be made in a studio or sound shielded room. With respect to the need of an excellent signal-to-noise-ratio, it is certainly not an option for noisy surroundings.

Re-speaking has advantages though. It makes it possible to adapt the spoken language for an audience with limited oral language proficiency. This would not be possible with automatic speech recognition.

Real-time speech-to-text conversion with speech recognition systems does not require special technical knowledge or training except for the fact that the SR- system has to be trained. For the user it is sufficient to speak correctly. However, linguistic knowledge and a kind of "thinking with punctuation" is necessary to dictate with punctuation marks.

Summary of speech recognition

Automatic speech recognition is not yet an option for speech-to-text transfer since phrase- and sentence boundaries are not recognized. However, speech recognition can be used for real-time speech-to-text conversion if a person re-speaks the original words. Re-speaking is primarily necessary for including punctuation and speaker identification but also for adapting the language to the language proficiency of the audience. Apart from an intensive and permanent training of the speech recognition engine, no special training is required. A sound-shielded environment is useful. The use of a speech recognition systems does not require any special training. Linguistic knowledge, however, is necessary for the chunking of the words and for adaptations of the wording.

3.2 Computer-assisted note taking (CAN)

With computer-assisted note taking (CAN), a person writes into an ordinary computer what a speaker says. However, as was discussed earlier, even professional writing speed is not sufficient to write down every word of a speech. To enhance writing speed, abbreviation systems are used in computer-assisted note taking which minimize the amount of key strokes per word. The note taking person types abbreviations or a mixture of abbreviations and long forms. An abbreviation-to-long-form dictionary translates the abbreviations immediately into the corresponding long form. On the screen, every word appears in its long form.

Realizations of CAN systems are widespread. On the one hand, small systems are incorporated in almost every word processing software. The so called "auto correction" translates given or self defined abbreviations into the corresponding long forms. On the other hand, there are very elaborated and well developed systems like e.g. C-Print which has been developed at the National Technical Institute for the DEAF at Rochester Institute of Technology (RIT 2005). This system uses phonetic rules to minimize the key strokes for every word. After a period of training with the system, the captionist is able to write with a higher speed. This allows for a high quality message transfer. However, the writing speed is still limited so that word-for-word transcripts are rather unusual, even with C-Print. With CAN-systems like C-Print, a message-to-message rather than a word-for-word transfer is produced.

The efficiency of CAN systems is mainly determined by the quality of the dictionary which translates the short forms into the corresponding long forms. The better the dictionary, the higher the typing speed potential.

Individually made dictionaries are mostly a collection of abbreviations like 'hv' for 'have' and 'hvt' for 'have to' etc. However, this kind of dictionary is limited insofar as the user has to know every abbreviation. Consequently, the amount of time which is needed for people to learn and to prevent them from forgetting the abbreviations once learned increases with the increase in the size of the dictionary.

Elaborated systems like C-Print use rule-based short-to-long translations. Here, the captionist has to learn the rules of transcription. One rule could be that only consonants but not vowels are written down. The resulting ambiguities (e.g. 'hs' for 'house' and 'his') have to be resolved by a second rule. However, orthographic transcription rules turned out to be rather complicated – at least in English. Therefore, systems like C-Print are often based on a set of rules which are in turn based on a phonetic transcription of the spoken words. On the basis of a set of shortening rules, the note taking person does not write certain graphemes but phonemes of the spoken words.

Summary of CAN-systems:

CAN-systems can be used for real-time speech-to-text conversion if a message-to-message transfer is sufficient. For word-for-word transfers, the typing speed of CAN-systems is not high enough.
The quality and speed of the transfer depends on the kind and quality of the dictionary which translates abbreviations or shortened words into the corresponding readable long forms. To use a CAN-system, the note taking person needs to learn either the abbreviations of the short-to-long dictionary or the rules of short-phoneme/grapheme-to-long-grapheme conversion the dictionary is based on.

Linguistic knowledge is necessary for adaptations of the wording.

3.3 Communication access real-time translation (CART)

Communication access real-time translation (CART) uses stenography in combination with a computer based dictionary. The phonemes of a word are typed on a steno keyboard which allows the coding of more than one phoneme at a time. It is thus possible to code e.g. one syllable by a simultaneous key press with up to all 10 fingers: The left keys on the keyboard are used to code the initial sound of the syllable, the down keys code the middle sound and the right keys of the keyboard code the final sound of the syllable. For high frequency words or phrases, prefixes and suffixes, abbreviations are used.

The phonetic code of the words or the respective abbreviation is immediately translated into the corresponding long form by a sophisticated dictionary. An example (taken from www.stenocom.de, cf. Seyring 2005) can illustrate the advantage with respect to typing speed:

a) typing on a normal keyboard: 88 strokes
Ladies and Gentlemen! The people want to have calculability and stability.

b) Same words in machine steno code: 12 strokes
(The code between two spaces is 1 stroke, typed with up to 10 fingers.)
*HRAEUPLBG STPH T PAOEPL WAPBT TO*F KAL KUL BLT APBD STABLT FPLT*

The parallel typing with CART systems results in a high typing speed which is sufficient for word-for-word transcripts in real-time. The phonetic transcription reduces ambiguities between words and allows real-time accuracy levels of more than 95%. Moreover, if the audience is not interested in word-for-word conversion, CART systems can also be used for message-to-message transfers since they allow adaptations of the wording in real-time.

CART-systems can be used in silent or noisy surroundings, their efficiency mainly relies on the education of the person who does the writing. However, the education of the speech-to-text provider is one of the most limiting factors of CART systems. 3-4 years of intensive education with a lot of practicing are the minimum for a person to become a CART speech-to-text provider who produces text in sufficient quality (less then 4% of errors) and speed (ca. 150 words per minute). The second limitation of CART is the costs for the steno system of around 10.000 Euro.

Summary of CART-systems:

CART systems are highly flexible tools for real-time speech-to-text conversion. They can be used in noisy or silent surroundings for word-for-word as well as for message-to-message transfer. The limitations of CART are located outside of the system, i.e.

- the long period of training which is needed to become a good CART provider
- the costs of the steno system

3.4 Comparison of Speech Recognition, CAN- and CART-systems

	Speech Recognition with re-speaking	**Computer-Assisted Note-taking**	**Communication Access Real-time Translation**
Exact word protocols	Yes	almost, but needs a lot of training and a sophisticated dictionary	Yes
Language adaptations	Possible with re-speaking	Yes	Yes
Education to use the method	Some hours for initial training of SR-system	some weeks- months	3-4 years
Special conditions	Minimum background noise	None	None
Cost of equipment[i]	100-200 € SR-system 50-100 € good microphone (opt.) 1.000 Euro notebook	1.000 € notebook (+ licence for the dictionary)	~ 10.000 € steno machine 1.000 € notebook (+ licence for the steno-longhand dictionary)

Table 1: Speech recognition, computer-assisted note-taking and communication access real-time translation in comparison.

4 Text adaptation

Spoken and written forms of language rely on different mechanisms to transfer messages. Speech for instance is less grammatical and less chunked than text. A real-time speech-to-text conversion - even if it is a word-for-word service - has to chunk the continuous stream of spoken words into sentences and phrases with respect to punctuation and paragraphs in order for the text to be comprehensible. A correction of grammatical slips might be necessary, too, for word-for-word conversions and even more corrections my be necessary for an audience with less language proficiency. While intonation may alleviate incongruencies in spoken language, congruency errors easily cause misinterpretation in reading.

The transfer from spoken into written language patterns is only one method of text adaptation. As discussed earlier, the speech-to-text provider might also be asked to adapt the written text to the language proficiency of the audience. Here, the challenge of word-for-word transfer shifts to the challenge of message transfer with a reduced set of language material. A less skilled audience might be overstrained especially with complex syntactical structures and low frequent words and phrases. The speech-to-text provider therefore needs to know whether a word or phrase can be well understood or should better be exchanged with some more frequent equivalents. S/he also has to know how to split long and complex sentences into simpler structures to make them easier to understand.

The know how of text adaptation with respect to the needs of the audience is highly language- and field-specific. People who become C-Print captionists learn to use text condensing strategies which is mainly aimed at reducing key strokes (RIT 2005) but might also reduce grammatical complexity and lexical problems. However, a recent study on the effects of summarizing texts for subtitling revealed that “summarizing affects coherence relations, making them less explicit and altering the implied meaning” (Schilperoord et al. 2005, p.1). Further research has to show whether and how spoken language can be condensed in real-time without affecting semantic and pragmatic information.

For German, it has already been shown that test questions can (offline) be adapted linguistically without affecting the content of the question. That is, many words and structures can be replaced by equivalents that are easier to understand (cf. Cremer 1996; Schulte 1993; Wagner et al. 2004). Further research will have to show whether this kind of text adaptation on word-, sentence- and text level (in German called "Textoptimierung") can also be realized in real-time.

5 Presentation format

The last challenge of real-time speech-to-text transfer is the presentation of the text on the screen in a way that reading is optimally supported. The need to think about the presentation format is given as the text on the screen is moving which is a problem for the reading process. We usually read a fixed text, and our eyes are trained to move in saccades (rapid eye movements) on the basis of a kind of preview calculation with respect to the next words (cf. Sereno et al. 1998). But in real-time speech-to-text systems, the text appears consecutively on the screen and new text replaces older text when the screen is filled. A word-by-word presentation as a consequence of word-for-word transcription could result in less precise saccades which subsequently decreases the reading speed. Reading might be less hampered by a presentation line-by-line, as it is e.g. used in C-Print (cf. the online presentation at http://www.rit.edu/~techsym/detail.html#T11C). However, for slower readers, also line-by-line presentation might be problematic since the whole "old" text is moving upwards whenever a new line is presented. As a consequence, the word which was actually fixated by the eyes moves out of the fovea and becomes unreadable. The eyes have to look for the word and restart reading it.

The optimal presentation of real-time text for as many potential readers as possible is an issue which is worth further research, not only from the perspective of real-time transcription but also for subtitling purposes.

6 Perspectives

Real-time speech-to-text transfer is already a powerful tool which provides people with a hearing impairment access to oral communication. However, elaborated dictionaries as they are needed for efficient CAN- or CART-systems are not yet developed for many languages. Without those dictionaries, the systems can not be used.

Linguistic research has to find easy but efficient strategies for the real-time adaptation of the wording in order to make a message understandable also for an audience with limited language proficiency.

Finally, the optimal presentation of moving text to an audience with diverging reading abilities is a fascinating research field not only for real-time speech-to-text services but with respect to the presentation of movable text in general.

7 References

Cremer, Inge (1996): "Prüfungstexte verstehbar gestalten". *Hörgeschädigtenpädagogik* 4, 50. Jahrgang, Sonderdruck.

Eugeni, Carlo(forthcoming): "Respeaking". To be presented at the MuTra Conference 'LSP Translation Scenarios', 30 April – 4 May 2007, Vienna (to be published in the Proceedings 2007).

Leitch, David & MacMillan, Trish (2002): "How Students With Disabilities Respond to Speech Recognition Technology in the University Classroom - Final Research Report on the Liberated Learning Project". http://www.liberatedlearning.com/research/FINAL%20YEAR%20III%20LLP%20REPORT.pdf, visited: 23.08.2005.

Perfetti, Charles & Sandak, Rebecca (2000): "Reading Optimally Builds on Spoken Language: Implications for Deaf Readers". *Journal of Deaf Studies and Deaf Education* 5(1). Winter 2000. 32-50.

Rochester Institute of Technology, National Technical Institute for the Deaf (2005): "C-Print Speech-To-Text-system". http://www.ntid.rit.edu/cprint/. visited: July, 21, 2005.

Schilperoord, Joost & de Groot, Vanja & van Son, Nic (2005): "Nonverbatim Captioning in Dutch Television Programs: A Text Linguistic Approach". *Journal of Deaf Studies and Deaf Education* 10(4). Fall 2005. 402-416.

Schlenker-Schulte, Christa (1991): *Konjunktionale Anschlüsse. Untersuchungsergebnisse zu Grundelementen kommunikativ-sprachlichen Handelns bei hörgeschädigten und hörenden Jugendlichen.* Reihe: Wissenschaftliche Beiträge aus Forschung, Lehre und Praxis zur Rehabilitation von Menschen mit Behinderungen (WB XXXVII). Villingen-Schwenningen: Neckar-Verlag.

Schulte, Klaus (1993): *Fragen in Fachunterricht, Ausbildung, Prüfung*. Villingen-Schwenningen: Neckar-Verlag.

Sereno, Sara C. & Rayner, Keith & Posner, Michael I. (1998): "Establishing a time-line of word recognition: evidence from eye movements and event-related potentials". *Neuroreport* 9(10). 2195-2200.

Seyring, Heidrun (2005): "Computer-compatible stenography". http://www.stenocom.de/english/system.htm, visited: 23.08.29005.

Stinson, Michael & Horn, Christy & Larson, Judy & Levitt, Harry & Stuckless, Ross (1999): "Real-Time Speech-to-Text Services". http://www.netac.rit.edu/publication/taskforce/realtime. last visit: 23.08.2005.

Stuckless, Ross (1999): "Recognition Means More Than Just Getting the Words Right". *Speech Technology* Oct/Nov 1999, 30. http://www.speechtechmag.com/issues/4_6/cover/381-1.html. visited: July, 21, 2005.

Wagner, Susanne & Kämpf de Salazar, Christiane (2004): "Einfache Texte – Grundlage für barrierefreie Kommunikation". In Schlenker-Schulte, Christa (ed.): *Barrierefreie Information und Kommuniaktion: Hören - Sehen – Verstehen in Arbeit und Alltag*. WBL. Villingen-Schwenningen: Neckar-Verlag.

Wagner, Susanne & Prinz, Ronald & Bierstedt, Christoph & Brodowsky, Walter & Schlenker-Schulte, Christa (2004): „Accessible Multimedia: status-quo, trends and visions". *IT - Information Technology 6*. 346-352.