

Hearing at Home – Communication support in home environments for hearing impaired persons

Jonas Beskow¹, Björn Granström¹, Peter Nordqvist¹, Samer Al Moubayed¹,
Giampiero Salvi¹, Tobias Herzke², Arne Schulz³

¹KTH Centre for Speech Technology, Stockholm, Sweden,
²HörTech, Oldenburg, Germany, ³OFFIS, Oldenburg, Germany

{beskow, bjorn, nordq, giampi}@speech.kth.se, sameram@kth.se, t.herzke@hoertech.de, arne.schulz@offis.de

Abstract

The Hearing at Home (HaH) project focuses on the needs of hearing-impaired people in home environments. The project is researching and developing an innovative media-center solution for hearing support, with several integrated features that support perception of speech and audio, such as individual loudness amplification, noise reduction, audio classification and event detection, and the possibility to display an animated talking head providing real-time speechreading support. In this paper we provide a brief project overview and then describe some recent results related to the audio classifier and the talking head. As the talking head expects clean speech input, an audio classifier has been developed for the task of classifying audio signals as *clean speech*, *speech in noise* or *other*. The mean accuracy of the classifier was 82%. The talking head (based on technology from the SynFace project) has been adapted for German, and a small speech-in-noise intelligibility experiment was conducted where sentence recognition rates increased from 3% to 17% when the talking head was present.

Index Terms: Sound classification, speech processing, hearing impairment, communication support, talking heads, speechreading

1. Introduction

Speech and sounds are important sources of information in our everyday lives for communication with our environment, be it interacting with fellow humans or directing our attention to technical devices with sound signals. For hearing impaired persons this acoustic information must be enhanced, supplemented or even replaced by cues using other senses. Even if vibrators could be used, we believe that the most natural modality to use is the visual, since speech is fundamentally audiovisual and these two modalities are complementary. We are hence exploring how different visualization methods for speech and audio signals may support hearing impaired persons. The goal in this line of research is to allow the growing number of hearing impaired persons equal participation in communication.

2. Hearing at home

The methods and solutions presented in this article are implemented, tested, and evaluated in an ongoing EU project named Hearing at Home (HaH). The goal of the project is to develop the next generation of assistive devices that will allow the growing number of hearing impaired persons -

which predominantly includes the elderly - equal participation in communication and empowers them to play a full role in society. The project focuses on the needs of hearing impaired persons in home environments. Formerly separated devices like personal computer, Hi-Fi system, TV, digital camera, telephone, fax, intercom and services like internet access, VoIP, Personal Information Management (PIM), pay TV and home automation grow together, often to be accessible via a TV set connected to a PC or set top box (STB) that implements interfaces to network gateways as well as to home automation services. The TV becomes the central Home Information and Communication (HIC) platform of the household in the communication society.

To address hearing-impaired users, the HIC platform processes the audio output signals to improve speech intelligibility.

The system classifies of the acoustical environment present in the TV signal, applies appropriate noise reduction algorithms, and attenuates overly loud commercial breaks. Additionally a multi-band dynamic compression as it is typically done in hearing aids is applied to ensure best audibility for the hearing impaired user. All audio processing is performed within the HörTech “Master Hearing Aid” [1] software framework.

To further increase speech intelligibility, an artificial head can be displayed that performs lip movements controlled by the audio signal.

This article first describes the audio processing to detect alarm signals, classify the background sound environment and perform the phoneme recognition of incoming speech. We then propose different visualization techniques and exemplify with a number of visualization applications for hearing impaired persons.

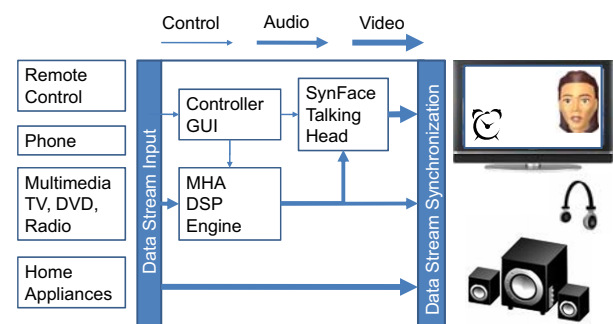


Figure 1: Overview of the Home Information and Communication platform (HIC) developed in the Hearing at Home project.

3. Sound classification

3.1. Speech and audio pre-processing

The audio in the system can be delivered to the user either by loudspeakers or by a headset. The MHA DSP engine is used to pre-process the audio to better fit the individual hearing loss. The purpose of the pre-processing is to increase the intelligibility and the listening comfort for the user. Noise reduction and sound classification are implemented in the MHA DSP engine. The following sections present results for the sound environment classification and the signal event detection.

3.1.1. Sound classification

Automatic sound classification has become an important area in various types of applications, for example surveillance systems, hearing aids, home automation, and communication support systems. The typical usage of a classification system is to support other functions, e.g. switching on/off the noise reduction in a hearing instrument.

This implementation uses two types of classifiers: sound environment classification and signal event detection. Sound environment classification is the process of detecting the overall acoustic situation, e.g. music, babble noise, traffic noise, and speech. Signal event detection is used when the task is to detect a particular signal that is generated from a well defined source. These types of signals are defined as signal events since they have a relatively short duration and are used for a specific purpose, e.g. the door bell signal.

3.1.2. Sound Environment Classification

The speech perception support functionality presented in this paper is only meaningful and should only be activated when the audio input is speech or speech in noise. A fully automatic system must therefore include at least a three category sound classifier that labels the incoming sound into three categories: clean speech, speech in noise and other.

The general principle for sound classification is presented in a separate work [2]. A sound environment classifier based on hidden Markov models and delta-cepstrum features has also been investigated [3].

The solution presented here uses three state left-right hidden Markov models. The training and evaluation material was collected from various TV shows and from real-life recordings. Several features were used to extract information from the audio signal, 'Onset Strength', 'Tonality', 'Modulation 16-64Hz', '2nd cepstrum', and 'Zero crossings'. The features are described in detail in a separate work [4]. Seventy percent of the material was used for training and the rest for evaluation. Table 1 presents the result of the classification. Perfect discrimination is not possible since the classes overlap, e.g. "other" includes babble noise.

Table 1. Confusion matrix, sound environment classification. Average hit rate 82%.

	1	2	3
1 Clean Speech	0.85	0.04	0.11
2 Other	0.02	0.88	0.10
3 Speech in Noise	0.04	0.24	0.72

3.1.3. Signal Event Detection

The main challenge of signal event detection in home environments is that the system must be robust against background noises, e.g. vacuum cleaning and music. In the approach presented here it is assumed that at least one spectral peak from the signal is visible above the noise spectrum. The frequency and the duration of the spectrum peaks for each signal event category are stored and compared against the current incoming sound.

The signal event categories used in the results presented here were door bell, digital clock, mobile phone, and phone. The signal sources were placed in an apartment at four different locations: living room, TV room, bedroom, and in the hall. A microphone was connected to a standard PC and placed in the hall. The signal sources were recorded in quiet and the spectrum peak characteristics were stored and used as models for the detection algorithm. After the training of the system, the same signals were presented again but now together with background noise at various signal-to-noise ratios. The types of background noises were TV, music and home noise (e.g. vacuum cleaning or porcelain clattering). The result from the evaluation is presented in Figure 2.

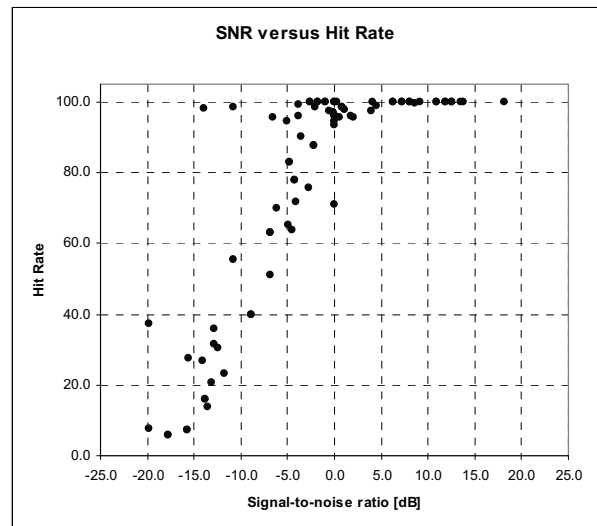


Figure 2: Acoustic signal event detection in noisy home environments. Hit rate as function of signal-to-noise ratio.

As expected there is a strong dependency between the hit rate (number of detected alarms over number of generated alarms) and the signal-to-noise ratio. The system works well with hit rates close to 100 percent for signal-to-noise ratios above 0 dB. The performance decreases rapidly when the signal-to-noise ratio is below 0 dB. It is interesting to notice that the hit rates for some of the signals are relatively high even at poor signal conditions, e.g. data points in the upper left of the graph. These results were achieved from mobile ring tones.

The false alarm rate is also an important design variable. It is defined as the number of unwanted detections, caused by other sound sources, per time unit. This number must be kept low in order for the user to accept the system. In this evaluation no false alarms were detected. Another solution for recognition of acoustical alarm signals for the profoundly deaf is using cepstrum features and hidden Markov models [5].

3.2. Visualization of signal events and sound environment

Visualization of the acoustic environment can be used to help hearing impaired persons to identify and minimize the uncertainty of the current listening situation. Illustrating the signal events and the listening environment is one way to increase the environment awareness for hearing impaired persons. A signal event could be illustrated with a symbol representing the action that caused the event. Similarly the current listening environment could be illustrated by e.g. one of three different symbols illustrating clean speech, speech in noise or babble, noise, and music. The usage of symbols for increasing the environment awareness will be further investigated and evaluated in the HaH-project.

4. SynFace visual speech support

For a hearing impaired person it is often necessary to be able to lip-read as well as hear the person they are talking with in order to communicate successfully. Often, only the audio signal is available, e.g. during telephone conversations or certain TV broadcasts. A visual speech support system, SynFace [6], is included in the HIC-platform. The idea behind SynFace is to try to re-create the visible articulation of the speaker, in the form of an animated talking head. The visual signal is presented in synchrony with the acoustic speech signal, which means that the user can benefit from the combined synchronized audiovisual perception of the original speech acoustics and the re-synthesized visible articulation. When compared to video telephony solutions, SynFace has the distinct advantage that only the user on the receiving end needs special equipment – the speaker at the other end can use any telephone terminal and technology – fixed, mobile or IP-telephony. In the HIC platform, SynFace can be applied not only to telephone signals, but to any audio stream, such as TV-audio, radio, CD/MP3 playback (audio-books) and so on, and can be switched on or off using the remote control.

4.1. Audio processing and animation

SynFace employs a specially developed real-time phoneme recognition system, based on a hybrid of recurrent neural networks (RNNs) and hidden Markov models (HMMs) [7], that delivers information regarding the speech signal to a speech animation module that renders the talking face to the computer screen using 3D graphics, as shown in Figure 3. The total delay from speech input to animation is only about 0.2 seconds, which is low enough not to disturb the flow of conversation. However, in order for face and voice to be perceived coherently, the acoustic signal also has to be delayed by the same amount. This delay is implemented in the HIC platform software.

The SynFace recogniser may be trained specifically for each language that the system should support. Under the EU-project SynFace, where the core technology was developed, support for Swedish, English and Dutch was developed. However, as the technique is based on phoneme recognition it can also be used with varying level of success with languages for which it has not been explicitly trained, depending on the degree of similarity between the articulatory spaces occupied by the languages in question.

In the HaH project, the set of supported languages has been extended to include German. The German recognizer was trained on 4000 speakers from the German SpeechDat database, and reached a frame level phoneme recognition accuracy of 67%.

The talking head model (shown in figure 3) includes face, tongue and teeth models, and is based on static 3D-wireframe meshes that are deformed by applying weighted transformations to their vertices [8], in turn described by high-level articulatory parameters. These parameters include jaw opening, lip rounding, bilabial occlusion and bilabial occlusion. A real-time articulatory control model [9] is responsible for driving the talking head's lip, jaw and tongue movements based on the phonetic input derived by the speech recogniser.

Additional parameters for the face control expressions, enabling the talking face to display emotions and non-linguistic cues, such as eyebrow raising, eye gaze, head nodding etc. The talking head may hence potentially provide perception support for higher level information, such as prosody or speaker mood, as well as for the actual phonemes uttered. Current ongoing work in HaH aims at identifying emphatic stress in the audio signal and supplementing this with visual cues.

The talking head engine has recently been ported to the windows mobile platform, allowing it to be displayed on PDAs and other mobile devices [10], see figure 3 right.



Figure 3: The talking head model, to the right shown running on a mobile device.

4.2. Evaluation

Previous publications have reported on the evaluation of the SynFace system and its sub components -- these evaluations are briefly summarized at the end of this section.

In a recent small experiment, we wanted to carry out a first stage evaluation of the newly developed German version of the SynFace system. A set of twenty short (4-6 words) sentences from the Göttinger satztest set [11], spoken by a male native German speaker, were presented to a group of six German listeners. The audio was processed using a 3-channel noise excited vocoder [12] to reduce intelligibility. 10 sentences were presented with audio only and 10 sentences were presented with SynFace support. 4 practice sentences were presented before the test started. The listeners were instructed to watch the screen and write down what they perceived.

The listeners responses were scored by counting the percentage of correctly perceived words for each of the two conditions. The mean score for the audio only condition was low, only 2.5%. With SynFace support, a mean score of 16.7% was obtained. While there was a large inter-subject variability in overall perception, subjects consistently showed an advantage for the SynFace condition. Figure 4 summarises the scores for each of the subject as well as the mean.

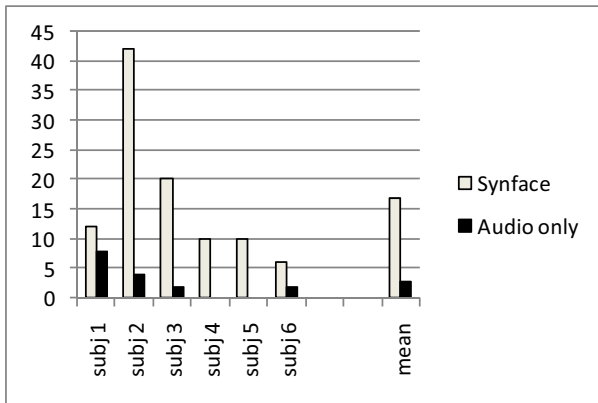


Figure 4: Early evaluation results for German version of SynFace (% correct word recognition)

The evaluations that took place during the SynFace project were of two kinds: audio visual intelligibility tests and user trials. The intelligibility tests [13] focused on measuring the intelligibility gain provided by the animated talking head by itself, assuming the existence of an ideal recognition system (simulated by forced alignment). These experiments, carried out for Swedish, English and Dutch, used normal-hearing subjects and sentences with degraded audio of the same kind that was used in the German test described above. Auditory signals were presented alone, with the synthetic face, and with a video of the original talker. Intelligibility on the purely auditory conditions was around 30% words correct, 50 % with talking head and close to 80% for the natural video. The magnitude of the intelligibility increase for the synthetic face compared to no face was broadly consistent, statistically reliable, and large enough to be important in everyday communication.

In the final user trials [14] of the prototype developed in the SynFace project, the system was evaluated by 49 users with varying degrees of hearing-impairment in UK and Sweden, in both lab and home environments, where they used SynFace to make phone calls. SynFace was found to give support to the users, especially in perceiving numbers and addresses and was considered an enjoyable way to communicate. A majority deemed SynFace to be a useful product.

5. Conclusions

This work presents solutions for individual parts included in the Hearing at Home project. An evaluation of the total benefit of the system is planned towards the end of this year. However, the intermediate results are promising both regarding the visual speech perception support, with a sixfold intelligibility increase over the audio-alone condition for the new German system, an average sound classification hit rate of 82% and signal event detection close to 100 percent for signal-to-noise ratios above 0 dB.

While the synthetic talking head has been shown to work well in telephony applications before, the Hearing at Home setting poses new challenges, due to the diversity of different audio streams available – TV-broadcasts, radio, film and music. The solution to this problem is a separate audio classification system that controls the further processing

steps. If the input is found to be too noisy, the talking face can be disabled altogether, because it is more likely to be of more harm than good. A less drastic option is to enable noise suppression algorithms available in the HIC platform to pre-filter the audio before it is fed to SynFace. This solution has yet to be evaluated.

6. Acknowledgements

The HaH project is funded by the EU (IST-045089). We would like to thank other project members at KTH, Sweden, HörTech, OFFIS, and ProSyst, Germany, VIATAAL, the Netherlands, and Telefonica I&D, Spain.

7. References

- [1] Grimm, G., & Herzke, T., & Berg, D., & Hohmann, V.. "The Master Hearing Aid: A PC-Based Platform for Algorithm Development and Evaluation". *Acta Acustica United with Acustica*, 92(4), 618-628, 2006
- [2] Nordqvist, P. "Sound Classification in Hearing Instruments," PhD Thesis, 2004
- [3] Nordqvist, P., and Leijon, A.. "An efficient robust sound classification algorithm for hearing aids," *J. Acoust. Soc. Am* 115, 3033-3041, 2004
- [4] Büchler, M. C. "Algorithms for sound classification in hearing instruments", PhD Thesis, 2002.
- [5] Oberle, S., and Kaelin, A. "Recognition of acoustical alarm signals for the profoundly deaf using hidden Markov models," in *IEEE International symposium on Circuits and Systems (Hong Kong)*, pp. 2285-2288., 1995
- [6] Beskow, J., Karlsson, I., Kewley, J., & Salvi, G. "SYNFACE - A talking head telephone for the hearing-impaired". In Miesenberger, K., Klaus, J., Zagler, W., & Burger, D. (Eds.), *Computers Helping People with Special Needs* (pp. 1178-1186). Springer-Verlag., 2004.
- [7] Salvi, G. "Dynamic behaviour of connectionist speech recognition with strong latency constraints". *Speech Communication*, 48(7), 802-818, 2006
- [8] Beskow, J. "Animation of Talking Agents", *Proceedings of AVSP'97*, pp. 149-152, 1997
- [9] Beskow, J. "Trainable articulatory control models for visual speech synthesis". *Journal of Speech Technology*, 4(7), 335-349, 2004.
- [10] Gjermani, T. Integration of an animated talking face model in a portable device for multimodal speech synthesis. Master of Science Thesis, KTH, Stockholm, Sweden, 2008
- [11] Wesselkamp, M., "Messung und Modellierung der Verständlichkeit von Sprache". Dissertation, Universität Göttingen, 1994
- [12] Shannon, R. V., Zeng, F-G., Kamath, V., Wygonski J. and Ekelid, M. "Speech recognition with primarily temporal cues", *Science*, 270, pp. 303-304, 1995
- [13] Siciliano, C., Williams, G., Beskow, J., Faulkner, A. "Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired". In: *Proc. of ICPhS.* . pp.131-134, 2003
- [14] Agelfors, E., Beskow, J., Karlsson, I., Kewley, J., Salvi, G., and Thomas, N.: "User Evaluation of the SYNFACE Talking Head Telephone". in K. Miesenberger et al. (Eds.): *ICCHP 2006, LNCS 4061*, pp. 579-586, 2006