INTERSPEECH 2009 BRIGHTON

Virtual Speech Reading Support for Hard of Hearing in a Domestic Multi-Media Setting

Samer Al Moubayed[1], *Jonas Beskow*[1], *Ann-Marie Öster*[1], *Giampiero Salvi*[1], *Björn Granström*[1], *Nic van Son*[2], *Ellen Ormel*[2]

[1] KTH Centre for Speech Technology, Stockholm, Sweden.
[2] Viataal, Nijmegen, The Netherlands.

sameram@kth.se,{beskow, annemarie, giampi, bjorn}@speech.kth.se, n.vson@viataal.nl, elleno@socsci.ru.nl

Abstract

In this paper we present recent results on the development of the SynFace lip synchronized talking head towards multi-linguality, varying signal conditions and noise robustness in the Hearing at Home project. We then describe the large scale hearing impaired user studies carried out for three languages. The user tests focus on measuring the gain in Speech Reception Threshold in Noise when using SynFace, and on measuring the effort scaling when using SynFace by hearing impaired people. Preliminary analysis of the results does not show significant gain in SRT or in effort scaling. But looking at inter-subject variability, it is clear that many subjects benefit from SynFace especially with speech with stereo babble noise.
Index Terms: hearing impairment, communication support, talking heads, visual speech reading, multilingual.

10.21437/Interspeech.2009-442

1. Introduction

There is a growing number of hearing impaired persons in the society today. In the ongoing EU-project Hearing at Home (HaH) [1], the goal is to develop the next generation of assistive devices that will allow this group - which predominantly includes the elderly – equal participation in communication and empower them to play a full role in society. The project focuses on the needs of hearing impaired persons in home environments.

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. One of the goals of the HaH project is to study the use of visual lip-reading support by hard of hearing people for home information, home entertainment, home automation, and home care applications running on a STB-like Home Information and Communication (HIC) platform.

2. Developments of the SynFace Phoneme Recogniser

SynFace [2] is a supportive technology for hearing impaired persons, which aims to re-create the visible articulation of a speaker, in the form of an animated talking head. SynFace employs a specially developed real-time phoneme recognition system, based on a hybrid of recurrent artificial neural networks (ANNs) and hidden Markov models (HMMs), that delivers information regarding the speech signal to a speech animation module that renders the talking face to the computer screen using 3D graphics.

Table 1: *Number of connections in the RNN and % correct frames of the SynFace RNN phonetic classifiers*

Language	Connections	% correct frames
Flemish	186,853	51.0
German	541,430	61.0
Swedish	541,250	54.2

SynFace previously has been trained on four languages: English, Flemish, German and Swedish. The training used the multilingual SpeechDat corpora. To align the corpora, the HTK (Hidden markov models ToolKit) based RefRec recogniser [2] was trained to derive the phonetic transcription of the corpus. Table 1 presents the correct frame rate of the recognizers of the languages used in the Hearing at Home project: Flemish, German and Swedish.

2.1. Varying Signal Conditions

There are a number of factors that influence the performance of the SynFace system when used in the hearing at home setting. As SynFace originally was developed for telephone usage, it was optimized for telephone bandwidth audio. In Hearing at Home, however, SynFace is expected to work with a variety of audio sources of varying quality, including wideband speech (16 kHz sampling frequency or better) and speech with varying signal level, or with various background noises (movie soundtracks etc). This section describes developments and experiments related to this situation.

In the HaH project, SynFace is employed in a range of applications that include speech signals that are streamed through different media (Telephone, Internet, TV). The signal is often of a higher quality compared to the land-line telephone settings. This opens the possibility for improvements in the signal processing part of the system.

In order to take advantage of the available audio band in these applications, the SynFace recogniser is trained on wide-band (16 kHz) speech data from the SpeeCon corpus [3]. SpeeCon contains recordings in several languages and conditions. Only recordings in office settings of Swedish were chosen.

The training of the wideband phoneme recognizer resulted in a 68% correct phoneme frame rate, which is a large improvement in phoneme recognition accuracy when compared to the other versions. In order to confirm that this improvement also transfers to increased intelligibility in the SynFace application, a small scale SRT intelligibility experiment was performed (see section 3 for description of the

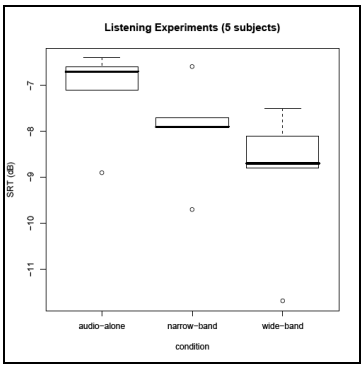

Figure 1: *Box plot of the Speech Reception Threshold (SRT dB – lower is better) for different conditions.*

method). The conditions include audio alone and SynFace driven by different versions of the recogniser. The tests were carried out using five normal hearing subjects. The stimuli consisted of lists of 5 words randomly selected from a set of 50 words. A training session was performed before the real test to control the learning effect. Figure 1 shows a box plot of the SRT levels obtained in the different conditions. An ANOVA analysis and successive multiple comparison analysis confirm that there is a significant decrease (improvement) of SRT ($p < 0.001$) for the wide-band recogniser over the narrow-band trained network and the audio-alone condition.

Previous evaluations of SynFace phoneme recognizer were done on speech signals with high Signal-to-Noise Ratio. These evaluations give good insight on the performance of SynFace in many signal conditions where clean speech is transmitted in a different channel than noise. Nevertheless, in other situations, separation of noise and speech is more difficult. To study the performance of the SynFace phoneme recognizer to deal with noisy speech, we evaluated the newly developed Swedish wideband recognizer on 4 different types of noises, namely: Stationary noise, 6-speakers stereo babble noise, stereo traffic noise and action movie noise, and with different SNR levels ranging from -20 to 30 dB. The correct frame rate of the recognizer is shown in Figure 2. The results show big degradation in the correct frame rate when these signals have a lower SNR than 20 dB. This result is expected since SynFace recognizer is not trained on noisy signals; from these results, it appears that using the current SynFace recognizers on noisy speech is impractical.

In an experiment to test the adaptation flexibility of the phoneme recognizer to noisy signals, the wideband Swedish phoneme recognizer (Neural Network) is trained on a small amount of noisy speech. Traffic noise was added to around 1 hour of speech recordings of the SpeeCon corpus, divided into 4 partitions with different SNRs (0, 4, 8 and 12). Figure 3 presents the difference in % correct phoneme frame rate between the original SynFace network and the adapted one on noisy speech. An interesting result is that the adapted network shows recognition enhancement, not only on the SNR range presented in the adaptation data, but generalized over all the SNR dimension [-20..+30] dB. This shows promising result on the ability of the SynFace recogniser to be adapted to specific noise types.

2.2. Multilinguality and Language Mapping

A factor that could limit the spread of the SynFace technology is the necessity to optimise the system for each new language. In principle, SynFace could be made language independent,

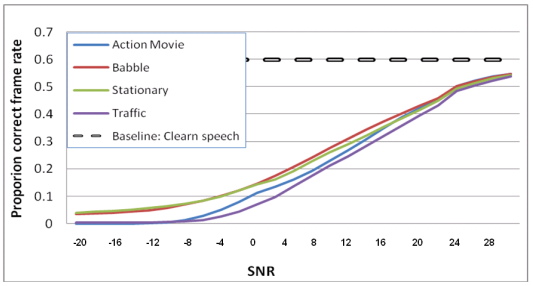

Figure 2*: Correct frame rate of the SynFace Swedish Wideband recognizer on noisy speech with different SNR levels*

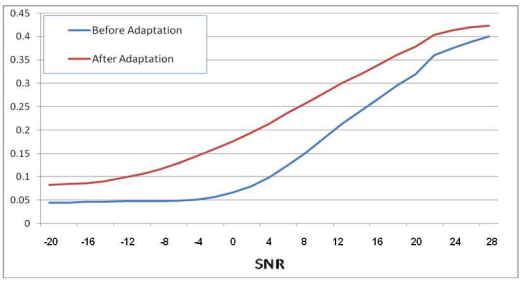

Figure 3: *The difference in correct frame rate of the SynFace recognizer before and after adaptation to noisy speech over different SNR levels.*

because it does not rely on lexical and grammatical information, and can be viewed as a problem of acoustic to articulatory inversion. However, inversion is still not feasible with the constraints imposed by the SynFace settings (real time, low latency, and low computational demands). We investigate the possibility of using models trained on a specific language to recognise phonemes in another language. This has the clear advantage of requiring smaller amounts of data and training time than the development of new language models.

SynFace is currently available in Swedish, German, Flemish and English. In order to investigate the possibility of using the current recognition models on new languages, we performed cross-language evaluation tests.

Each language has its unique phonetic inventory. In order to map between the inventory of the recognition model and that of the target language, we considered three different paradigms.

1. In the first case we rely on perfect matching of the phonemes. This is a very strict evaluation criterion because it does not take into account the acoustic and visual similarities between phonemes in different languages that do not share the same phonetic code.
2. The second mapping criterion considers as correct the association between model and target phonemes that was most frequently adopted by the recognition models on that particular target language.
3. The third paradigm is to take advantage of the posterior probabilities the phoneme recogniser generates for the input signal. In this way, we can in principle, obtain better results than the above upper bound, and possibly even better results than for the original language. In the experiments the probability mapping was performed by a one layer neural network that implements linear regression.

An experiment was done to map the English, Flemish and German recogniser output to Swedish language using 30 minutes of transcribed speech from the SpeeCon database [3].

The regression based approach (nr 3) outperformed the first two approaches (forced phoneme mapping, and best match mapping). The German recogniser mapped to Swedish gave 45% accuracy, which is close to that of the original Swedish trained recogniser (52% on the same test set).

3. User Studies

SynFace has been previously evaluated by subjects with many regards [4,5,6]. In the present study, a large scale test for the use of SynFace as an audio-visual support on hearing impaired people with different hearing loss levels. In particular, we wanted to evaluate SynFace with regard to two questions:

SRT (Speech Reception Threshold) measurement: Is SynFace useful at increasing the sentence intelligibility (% recognition of correct words), compared to the audio signal alone?

Effort scaling with/without SynFace: Is using SynFace, as a talking head, effortless, and easy to use by hearing impaired persons?

3.1. Method

SRT or Speech Reception Threshold is the speech signal SNR when the listener is able to understand 50% of the words in the signal. In this test, the SRT value is measured one time with a speech signal alone without SynFace, with two types of noise, and another time with the use of (when looking at) SynFace. If the SRT level has decreased when using SynFace, that means the subject has benefited from the use of SynFace, since the subject could understand 50% of the words with a higher noise level than when listening to the audio signal alone.

To calculate the SRT level, a recursive procedure described in [7] is used, where the subject listens to successive sets of 5 words, and depending on how many words the subject recognises correctly, the SNR level of the signal is changed so the subject can only understand 50% of the words.

The SRT value is estimated for each subject in five conditions, a first estimation is used as training to eliminate any training effect, and this was recommended in [7]. Two SRT values are estimated in the condition of speech signal without SynFace, but with two types of noise, Stationary noise, and Babble noise (with 6 speakers). The other two estimations are for the same types of noise, but with the use of SynFace, that is when the subject is looking at the screen at SynFace, and listening in the head-phones to a noisy signal.

In the effort scaling test, the easiness of using SynFace by hearing impaired persons was targeted. The test aims at measuring how difficult it is for the subjects to look at the SynFace talking head, while listening to speech signals. To establish this, the subject has to listen to sentences in the headphones, sometimes with looking at SynFace and sometimes without looking at SynFace, and choose using a touch screen, a value on a pseudo continuous scale, ranging from 1 to 6, telling how difficult it is to listen to the speech signal transmitted through the head-phones.

3.2. Subjects

The tests are performed on five groups of hearing impaired subjects with different hearing impairment levels (Mild, Moderate, Severe and Subjects with cochlear implants). Every group consists of 15 subjects. Table 2 shows information and the location of the user groups.

Table 2: *Description of the hearing impaired test subjects groups.*

	Swedish	German	Flemish
# Subjects	15	15 + 15	15 + 15
Hearing Impairment	Moderate	Mild + Moderate	Severe + Cochlear Implants
Location	KTH Sweden	Hörtech Germany	Viataal Netherland

3.3. Preliminary Analysis and Results

Mean results of the SRT measurement tests are presented in Figure 4. The figure shows the inherent level of SRT value for the different hearing impairment groups (cochlear implants with a noticeably higher level than the other groups), as well as the difference in SRT value with and without using SynFace. The mean values do not show significant decrease or increase in the SRT level when using SynFace than with audio-only conditions. Nevertheless, when looking at the performance of the subjects individually, a high inter-subject variability is clear which means that certain subjects have benefited from the use of SynFace. Figure 5 shows the sorted delta SRT value per subject for the Swedish moderate hearing impaired subject for speech signals with stationary and babble noise. In addition to the high variability among subjects, it is clear that, in the case of babble noise, most of the subjects show benefit (negative delta SRT).

Regarding the results of the effort scaling, subjects at all locations, do not show significant difference in scaling value between the condition of speech with and speech without SynFace. But again, the scaling value shows a high inter-subject variability.

Another investigation we carried out was to study the effect of the SRT measurement lists length on the SRT value. As mentioned before, the SRT measurement used contained 20 lists of words, where every list contained 5 words, and one training measurement was done at the beginning to eliminate any training effect. Still, when looking at the average trend of the SRT value over time for each list, the SRT value was decreasing. This can be explained as an ongoing training throughout the lists for each subject. But when looking at the individual SRT value per test calculated after the 10th and the 20st list for each measurement, a surprising observation was

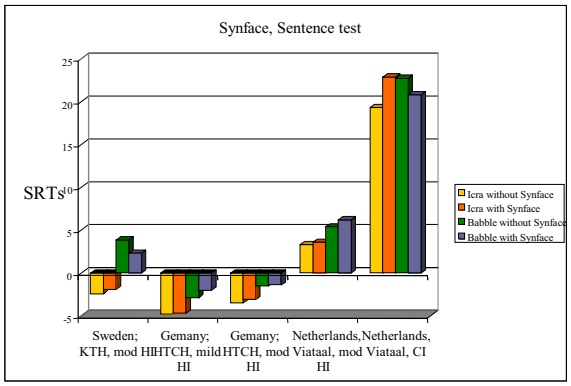

Figure 4: *Mean Scores on SRT measurement sentence test with Stationary and Babble noise condition With and Without Synface for Different Groups*

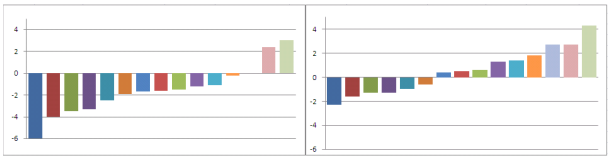

Figure 5: *The delta SRT value per subject for the Swedish moderate hearing impaired group (with SynFace-Without SynFace). Left: in the condition of speech with babble noise. Right: in the condition of speech with stationary noise.*

that for some of the measurements, the SRT value of the same measurement increased at the 20st list compared to the 10th list. Figure 6 presents the difference of SRT value at the 20st list and the 10th list for 40 SRT measurements which shows that although most of the measurements had a decreasing SRT value, some of them had an increasing one. This means that the longer the measurement is not always better (decreasing the learning effect).We suspect here that this can be a result of that the 20-lists long measurements are too long for the hearing impaired subjects, and that they might be getting tired and loosing concentration when the measurement is as long as 20 lists and hence requiring a higher SNR.

3.4. Discussion

Overall, the preliminary analysis of the results of both the SRT test and the effort scaling showed limited beneficial effects for Synface. However, the Swedish participants showed an overall beneficial effect for the use of Synface in the SRT test when listening to speech with babble noise.

Another possible approach when examining the benefit of using Synface may be looking at individual results as opposed to group means. The data shows that some people benefit from the exposure to Synface. In the ongoing analysis of the tests, we will try to see if there are correlations in the results for different tests per subject, and hence to study whether there are certain feature which characterize subjects who show consistent benefit from SynFace throughout all the tests.

In future work, it may be useful to include additional measurements, such as cognitive tasks which are related to speech reading, in order to gain more insight in what characterizes people who benefit from the use of Synface.

4. Conclusions

This paper summarizes the work done in developing, enhancing, adapting and evaluating the SynFace technology in Hearing at Home. SynFace, which originally was developed only for telephony application, has in the Hearing at Home project been developed and extended in several ways. Language support has been widened, not only by developing a German version of the system, but also by researching into methods of adding support for new languages in cost-effective ways, by using existing phonetic recognizers and mapping the output of these into new language sets, which appears to be a promising technique, that allows nearly the performance of a system trained on a full database (thousands of speakers) with only 30 minutes of speech from the new language.

The Hearing at Home project places the SynFace technology in a new context that also imposes new challenges. One is that SynFace is expected to work in a wide diversity of acoustic signal conditions, not only telephone speech as was the case previously. SynFace has to be able to take advantage of

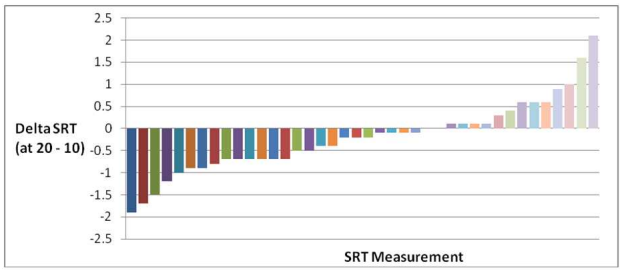

Figure 6: *Delta between the SRT estimated at the 20th list and the 10th list displayed by measurement.*

broadband audio, which has led to new versions of SynFace with improved performance. Perceptual experiments show that the Wideband version of SynFace gives a significant benefit for the SynFace condition when compared to both Narrowband versions and to audio only. In addition, SynFace may be exposed to speech mixed with various types of noises (TV or movie soundtracks etc.). Experiments indicate that SynFace performance is severely degraded by noises, but also that there is a potential for improvement when SynFace has been specially trained to deal with noisy speech. The paper also reports a preliminary analysis of the results from the user studies with hearing impaired subjects performed at three sites. On average, SynFace did not show a consistent advantage with the hearing impaired subjects, but there were some SynFace benefits, especially for speech-in-babble-noise, and for particular users.

5. 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.

6. References

[1] Beskow, J., Granström, B., Nordqvist, P., Al Moubayed, S., Salvi, G., Herzke, T., & Schulz, A. (2008). Hearing at Home – Communication support in home environments for hearing impaired persons. In Proceedings of Interspeech 2008. Brisbane, Australia.

[2] Lindberg, B., Johansen, F. T., Warakagoda, N., Lehtinen, G., Kai, Z., Gank, A., Elenius, K., & Salvi, G. (2000). A noise robust multilingual reference recogniser based on SpeechDat(II). In Proc of ICSLP 2000, 6th Intl Conf on Spoken Language Processing (pp. 370-373). Beijing.

[3] Iskra, D., Grosskopf, B., Marasek, K., Heuvel, H. V. D., Diehl, F., & Kiessling, A. (2002). Speecon - speech databases for consumer devices: Database specification and validation. In Proc. LREC, 2002 (pp. 329-333).

[4] Agelfors, E., Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K-E., & Öhman, T. (1998). Synthetic faces as a lipreading support. In Proceedings of ICSLP'98.

[5] Siciliano, C., Faulkner, A., & Williams, G. (2003). Lipreadability of a synthetic talking face in normal hearing and hearing-impaired listeners. In AVSP 2003-International Conference on Audio-Visual Speech Processing.

[6] Agelfors, E., Beskow, J., Karlsson, I., Kewley, J., Salvi, G., & Thomas, N. (2006). User Evaluation of the SYNFACE Talking Head Telephone. Lecture Notes in Computer Science, 4061, 579-586.

[7] Hagerman, B., & Kinnefors, C. (1995). Efficient adaptive methods for measuring speech reception threshold in quiet and in noise. Scand Audiol, 24, 71-77.