SynFace Phone Recognizer for Swedish Wideband and Narrowband Speech

Samer al Moubayed
KTH Centre for Speech Technology
Stockholm, Sweden
sameram@kth.se

Jonas Beskow
KTH Centre for Speech Technology
Stockholm, Sweden
beskow@kth.se

Giampiero Salvi
KTH Centre for Speech Technology
Stockholm, Sweden
giampi@kth.se

Abstract

In this paper, we present new results and comparisons of the real-time lips synchronized talking head SynFace on different Swedish databases and bandwidth. The work involves training SynFace on narrow-band telephone speech from the Swedish SpeechDat, and on the narrow-band and wide-band Speecon corpus. Auditory perceptual tests are getting established for SynFace as an audio visual hearing support for the hearing-impaired. Preliminary results show high recognition accuracy compared to other languages.

1 Introduction

For a hearing impaired person it is often necessary to be able to lip-read as well as hear the person they are listening to in order to communicate successfully. Often, however, only the audio signal is available. The idea behind SynFace (Beskow et al, 2004) is to try to re-create the visible articulation of the speaker, in the form of an animated talking head.

Originally, SynFace was developed for telephone conversation support. Presently, we are extending the scope of SynFace to other audio sources such as television- and radio broadcasts and audio books. This is done in the Hearing at Home (HaH) project (Beskow et al, 2008), which 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 that besides SynFace also includes features such as individual loudness amplification, noise reduction, audio classification and event detection.

The use of SynFace for non-telephone speech implies a potential for improvement, in that speech signals with a wider spectrum are available. Consequently, we are investigating the added benefit of training a recognizer to take advantage of wideband speech (16 kHz sampling frequency or higher).

2 SynFace and the Phone Recognizer

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) (Salvi, 2003), that delivers information regarding the speech signal to a speech animation module that renders the talking face to the computer screen using 3D graphics.

In this work, the ANN structure used in SynFace is a three layers network with a 400 neurons recurrent hidden layer.

3 Swedish Narrowband Recognizer

Previously, SynFace was trained on the Swedish SpeechDat corpus (Elenius et al, 1997), The corpus used for training contains 8 KHz telephone speech, consisting of 5000 speakers' speech. The corpus was aligned using an HMM phone level aligner using word level transcription. Table 1 presents the specification and the training accuracy for the NB Swedish SpeechDat.

4 Swedish Wideband Recognizer

To train SynFace recognizer for a wide band data, the Speecon corpus is used (Iskra et al, 2002). Speecon contains several languages and speech conditions recordings. Only office recordings of Swedish were chosen. The corpus contains word level transcriptions, and annotations for speaker noise, background noise, and filled pauses. To balance the number of frames in the corpus per phone, the silence at the boundaries of every utterance were cut. For aligning the

corpus, the NALIGN aligner (Sjölander, 2003) is used, using HMM models for wide band Swedish speech. Table 1 presents the specification of the training data, and the obtained frame level accuracy after training the neural network.

Table 1: Summary of databases used for training of the NB and WB recognizers, and resulting network performance

Corpus name	SpeechDat (NB)	Speecon (WB)
Frames number	70 M	~15 M
Sampling rate	8KHz	16KHz
Speakers	5000	550
Training time	~2.5 Month	4 weeks
Accuracy	54.2%	62.9%

These results show that although the Narrowband network was trained on a significantly larger and more varied dataset, the Wideband network obtained higher accuracy; this can be of different reasons, like spectral resolution, data consistency, the speech SNR.

5 Evaluation - Work in progress

This abstract represents a work in progress; more detailed evaluation results will be given at the time of the conference. Presently, we only present raw frame-level accuracy of the ANNs. In order to better evaluate the performance of the different recognizers we will also:

- Train a network on a narrowband speech using the same data for wideband but after filtering to telephone bandwidth and down-sampling to 8 KHz, in order to have a more constrained comparison between the narrowband and the wideband networks for Swedish.
- Study differences in phone level recognition between the different networks in terms of confusion matrices to gain insight into what confusions are to be expected due to bandwidth reduction, and how these map to the viseme classes used in the facial animation rendering
- Perform a small scale intelligibility experiment on human subjects based on an adaptive speech reception threshold paradigm (Hagerman and Kinnefors, 1995), in order to find the SNR levels at which equal (50%) word recognition rates are achieved for different SynFace conditions (different ANNs) and for audio alone.

Acknowledgments
This work is carried out under the Hearing at Home project (EU IST-045089).

References

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.

Beskow, J., Karlsson, I., Kewley, J., & Salvi, G. (2004). 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.

Iskra, D., Grosskopf, B., Marasek, K., Van Den Heuvel, H., Diehl, F., and Kiessling, A. (2002) . Speecon - speech databases for consumer devices: Database specification and validation

Elenius, K., & Lindberg, J. (1997). SpeechDat - Speech databases for creation of voice driven teleservices. In Bannert, R., Heldner, M., Sullivan, K., & Wretling, P. (Eds.), Proc of Fonetik -97, Dept of Phonetics, Phonum 4 (pp. 61-64). Lövånger/Umeå.

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

Salvi, G. (2003). Truncation error and dynamics in very low latency phonetic recognition. In Non-linear Speech Signal Processing, NOLISP2003

Sjölander, K. (2003). An HMM-based system for automatic segmentation and alignment of speech. In Proc of Fonetik 2003, Umeå University, Dept of Philosophy and Linguistics PHONUM 9 (pp. 93-96).