Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2009, Article ID 191940, 10 pages
doi:10.1155/2009/191940

Research Article

SynFace—Speech-Driven Facial Animation for Virtual Speech-Reading Support

Giampiero Salvi, Jonas Beskow, Samer Al Moubayed, and Björn Granström

KTH, School of Computer Science and Communication, Deptarment for Speech, Music, and Hearing, Lindstedtsvägen 24, SE-100 44 Stockholm, Sweden

Correspondence should be addressed to Giampiero Salvi, giampi@kth.se

Received 13 March 2009; Revised 23 July 2009; Accepted 23 September 2009

Recommended by Gérard Bailly

This paper describes SynFace, a supportive technology that aims at enhancing audio-based spoken communication in adverse acoustic conditions by providing the missing visual information in the form of an animated talking head. Firstly, we describe the system architecture, consisting of a 3D animated face model controlled from the speech input by a specifically optimised phonetic recogniser. Secondly, we report on speech intelligibility experiments with focus on multilinguality and robustness to audio quality. The system, already available for Swedish, English, and Flemish, was optimised for German and for Swedish wide-band speech quality available in TV, radio, and Internet communication. Lastly, the paper covers experiments with nonverbal motions driven from the speech signal. It is shown that turn-taking gestures can be used to affect the flow of human-human dialogues. We have focused specifically on two categories of cues that may be extracted from the acoustic signal: prominence/emphasis and interactional cues (turn-taking/back-channelling).

1. Introduction

For a hearing impaired person, and for a normal hearing person in adverse acoustic conditions, 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. Apart from the lip movements, nonverbal visual information is also essential to keep a normal flow of conversation. Often, only the audio signal is available, for example, during telephone conversations or certain TV broadcasts. The idea behind SynFace is to try to recreate 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 synchronised audiovisual perception of the original speech acoustics and the resynthesised 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, mobilem, or IP-telephony.

Several methods have been proposed to drive the lip movements of an avatar from the acoustic speech signal with varying synthesis models and acoustic-to-visual maps. Tamura et al. [1] used hidden Markov models (HMMs) that are trained on parameters that represent both auditory and visual speech features. Similarly, Nakamura and Yamamoto [2] propose to estimate the audio-visual joint probability using HMMs. Wen et al. [3] extract the visual information from the output of a formant analyser. Al Moubayed et al. [4] map from the lattice output of a phonetic recogniser to texture parameters using neural networks. Hofer et al. [5] used trajectory hidden Markov models to predict visual speech parameters from an observed sequence.

Most existing approaches to acoustic-to-visual speech mapping can be categorised as either regression based or classification based. Regression-based systems try to map features of the incoming sounds into continuously varying articulatory (or visual) parameters. Classification-based systems, such as SynFace, consider an intermediate phonetic level, thus solving a classification problem, and generating the final face parameters with a rule-based system. This

(a) (b)

Figure 1: One of the talking head models used in SynFace, to the right running on a mobile device.

approach has proved to be more appropriate when the focus is on a real-life application, where additional requirements are to be met, for example, speaker independence, and low-latency. Ohman and Salvi [6] compared two examples of the two paradigms. A time-delayed neural network was used to estimate the face parameter trajectories from spectral features of speech, whereas an HMM phoneme recogniser was used to extract the phonetic information needed to drive the rule-based visual synthesis system. Although the results are dependent on our implementation, we observed that the first method could learn the general trend of the parameter trajectories, but was not accurate enough to provide useful visual information. The same is also observed in Hofer et al. [5] and Massaro et al. [7]. (Although some speech-reading support was obtained for isolated words from a single speaker in Massaro's paper, this result did not generalise well to extemporaneous speech from different speakers (which is indeed one of the goals of SynFace).) The second method resulted in large errors in the trajectories in case of misrecognition, but provided, in general, more reliable results.

As for the actual talking head image synthesis, this can be produced using a variety of techniques, typically based on manipulation of video images [8, 9] parametrically deformable models of the human face and/or speech organs [10, 11] or as a combination thereof [12]. In our system we employ a deformable 3D model (see Section 2) for reasons of speed and simplicity.

This paper summarises the research that led to the development of the SynFace system and discusses a number of aspects involved in its development, along with novel experiments in multilinguality, dependency on the quality of the speech input, and extraction of nonverbal gestures from the acoustic signal.

The SynFace architecture is described for the first time as a whole in Section 2; Section 3 describes the additional nonverbal gestures. Experiments in German and with wide-band speech quality are described in Section 4. Finally, Section 5 discusses and concludes the paper.

2. SynFace Architecture

The processing chain in SynFace is illustrated in Figure 2. SynFace employs a specially developed real-time phoneme recognition system, that delivers information regarding the speech signal-to-a speech animation module that renders the talking face on the computer screen using 3D graphics. The total delay from speech input to animation is only about 200 milliseconds, which is low enough not to disturb the flow of conversation, (e.g., [13]). However, in order for face and voice to be perceived coherently, the acoustic signal also has to be delayed by the same amount [14].

2.1. Synthesis. The talking head model depicted in Figures 1 and 2 includes face, tongue, and teeth, and is based on static 3D-wireframe meshes that are deformed using direct parametrisation by applying weighted transformations to their vertices according to principles first introduced by Parke [15]. These transformations are in turn described by high-level articulatory parameters [16], such as jaw opening, lip rounding and bilabial occlusion. The talking head model is lightweight enough to allow it to run at interactive rates on a mobile device [17]. A real-time articulatory control model is responsible for driving the talking head's lip, jaw and tongue movements based on the phonetic input derived by the speech recogniser (see below) as well as other facial motion (nodding, eyebrow movements, gaze, etc.) further described in Section 3.

The control model is based on the rule-based look-ahead model proposed by Beskow [16], but modified for low-latency operation. In this model, each phoneme is assigned a target vector of articulatory control parameters. To allow the targets to be influenced by coarticulation, the target vector may be under-specified, that is, some parameter values can be left undefined. If a target is left undefined, the value is inferred from context using interpolation, followed by smoothing of the resulting trajectory. As an example, consider the lip rounding parameter in a V_1CCCV_2 utterance where V_1 is an unrounded vowel, CCC represents a consonant cluster and V_2 is a rounded vowel. According to the rules set, lip rounding would be unspecified for the consonants, leaving these targets to be determined from the vowel context by linear interpolation from the unrounded V_1, across the consonant cluster, to the rounded V_2.

To allow for low-latency operation, the look-ahead model has been modified by limiting the look-ahead time window (presently a value of 100 milliseconds is used) which means that no anticipatory coarticulation beyond this window will occur.

For comparison, the control model has also been evaluated against several data-driven schemes [18]. In these experiments, different models are implemented and trained to reproduce the articulatory patterns of a real speaker, based on a corpus of optical measurements. Two of the models, (Cohen-Massaro and Ohman) are based on coarticulation models from speech production theory and one uses artificial neural networks (ANNs). The different models were evaluated through a perceptual intelligibility experiment, where the data-driven models were compared against

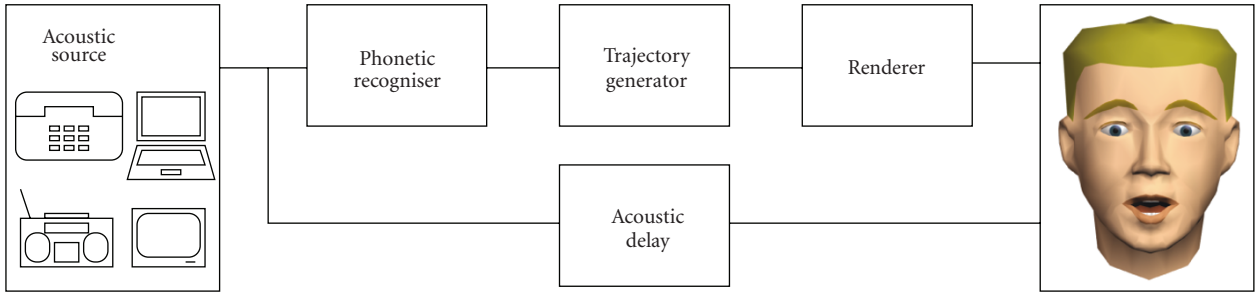

Figure 2: Illustration of the signal flow in the SynFace system.

Table 1: Summary of intelligibility test of visual speech synthesis control models, from Beskow [18].

Control model	% keywords correct
Audio only	62.7
Cohen-Massaro	74.8
Ohman	75.3
ANN	72.8
Rule-based	81.1

the rule-based model as well as an audio-alone condition. In order to only evaluate the control models, and not the recognition, the phonetic input to all models was generated using forced alignment Sjolander [19]. Also, since the intent was a general comparison of the relative merits of the control models, that is, not only for real time applications, no low-latency constraints were applied in this evaluation. This means that all models had access to all segments in each utterance, but in practise the models differ in their use of look-ahead information. The "Cohen-Massaro" model by design always uses all segments; the "Ohman" model looks ahead until the next upcoming vowel; while the ANN model, which was specially conceived for low-latency operation, used a constant look-ahead of 50 milliseconds.

Table 1 summarises the results; all models give significantly increased speech intelligibility over the audio-alone case, with the rule-based model yielding the highest intelligibility score. While the data-driven models seem to provide movements that are in some sense more naturalistic, the intelligibility is the single most important aspect of the animation in SynFace, which is why the rule-based model is used in the system.

2.2. Phoneme Recognition. The constraints imposed on the phoneme recogniser (PR) for this application are speaker independence, task independence and low latency. However, the demands on the PR performance are limited by the fact that some phonemes map to the same visemes (targets) for synthesis.

The phoneme recogniser used in SynFace, is based on a hybrid of recurrent neural networks (RNNs) and hidden Markov models (HMMs) [20]. Mel frequency cepstral coefficients (MFCCs) are extracted on 10 milliseconds spaced frames of speech samples. The neural networks are used to estimate the posterior probabilities of each phonetic class given a number of feature vectors in time [21]. The networks are trained using Back Propagation through time [22] with a cross-entropy error measure [23]. This ensures an approximately linear relation between the output activities of the RNN and the posterior probabilities of each phonetic class, given the input observation. As in Strom [24], a mixture of time delayed and recurrent connections is used. All the delays are positive, ensuring that no future context is used and thus reducing the total latency of the system at the cost of slightly lower recognition accuracy.

The posterior probabilities estimated by the RNN are fed into an HMM with the main purpose of smoothing the results. The model defines a simple loop of phonemes, where each phoneme is a left-to-right three-state HMM. A slightly modified Viterbi decoder is used to allow low-latency decoding. Differently from the RNN model, the decoder makes use of some future context (look-ahead). The amount of look-ahead is one of the parameters that can be controlled in the algorithm.

During the Synface project (IST-2001-33327), the recogniser was trained and evaluated on the SpeechDat recordings [25] for three languages: Swedish, English and Flemish. In Salvi [20, 26], the effect of limiting the look-ahead in the Viterbi decoder was studied. No improvements in the results were observed for look-ahead lengths greater than 100 milliseconds. In the SynFace system, the look-ahead length was further limited to 30 milliseconds, resulting in a relative 4% drop in performance in terms of correct frames.

3. Nonverbal Gestures

While enhancing speech perception through visible articulation has been the main focus of SynFace, recent work has been aimed at improving the overall communicative experience through nonarticulatory facial movements. It is well known that a large part of information transfer in face-to-face interaction is nonverbal, and it has been shown that speech intelligibility is also affected by nonverbal actions such as head movements [27]. However, while there is a clear correlation between the speech signal and the articulatory movements of the speaker that can be exploited for driving the face articulation, it is less clear how to provide meaningful nonarticulatory movements based solely on the acoustics. We have chosen to focus on two classes

of nonverbal movements that have found to play important roles in communication and that also may be driven by acoustic features that can be reliably estimated from speech. The first category is speech-related movements linked to emphasis or prominence, the second category is gestures related to interaction control in a dialogue situation. For the time being, we have not focused on expressiveness of the visual synthesis in terms of emotional content as in Cao et al. [28].

Hadar et al. [29] found that increased head movement activity co-occurs with speech, and Beskow et al. [30] found, by analysing facial motion for words in focal and nonfocal position, that *prominence* is manifested visually in *all* parts of the face, and that the particular realisation chosen is dependent on the context. In particular these results suggest that there is not *one* way of signalling prominence visually but it is likely that several cues are used interchangeably or in combination. One issue that we are currently working on is how to reliably extract prominence based on the audio signal alone, with the goal of driving movements in the talking head. In a recent experiment Al Moubayed et al. [4] it was shown that adding small eyebrow movements on syllables with large pitch movements, resulted in a significant intelligibility improvement over the articulation-only condition, but less so than a condition where manually labelled prominence was used to drive the gestures.

When people are engaged in face-to-face conversation, they take a great number of things into consideration in order to manage the flow of the interaction. We call this *interaction control*—the term is wider than turn-taking and does not presuppose the existence of "turns." Examples of features that play a part in interaction control include auditory cues such as pitch, intensity, pause and disfluency, hyper-articulation; visual cues such as gaze, facial expressions, gestures, and mouth movements (constituting the *regulators* category above) and cues like pragmatic, semantic, and syntactic completeness.

In order to investigate the effect of visual interaction control cues in a speech driven virtual talking head, we conducted an experiment with human-human interaction over a voice connection supplemented by the SynFace talking head at each end, where visual interaction control gestures were automatically controlled from the audio stream. The goal of the experiment was to find out to what degree subjects were affected by the interaction control cues. In what follows is a summary, for full details see Edlund and Beskow [31].

In the experiment, a bare minimum of gestures was implemented that can be said to represent a stylised version of the gaze behaviours observed by Kendon [32] and recent gaze-tracking experiments [33].

(i) A turn-taking/keeping gesture, where the avatar makes a slight turn of the head to the side in combination with shifting the gaze away a little, signalling a wish to take or keep the floor.

(ii) A turn-yielding/listening gesture, where the avatar looks straight forward, at the subject, with slightly raised eyebrows, signalling attention and willingness to listen.

(iii) A feedback/agreement gesture, consisting of a small nod. In the experiment described here, this gesture is never used alone, but is added at the end of the listening gesture to add to its responsiveness. In the following, simply assume it is present in the turn yielding/listening gesture.

The audio-signal from each participant was processed by a voice activity detector (VAD). The VAD reports a change to the SPEECH state each time it detected a certain number of consecutive speech frames whilst in the SILENCE state, and vice-versa. Based on these state transitions, gestures were triggered in the respective SynFace avatar.

To be able to assess the degree to which subjects were influenced by the gestures, the avatar on each side could work in one of two modes: ACTIVE or PASSIVE. In the ACTIVE mode, gestures were chosen as to encourage one party to take and keep turns, while PASSIVE mode implied the opposite—to discourage the user to speak. In order to collect balanced data of the two participants behaviour, the modes were shifted regularly (every 10 turns), but they were always complementary—ACTIVE on one side and PASSIVE on the other. The number 10 was chosen to be small enough to make sure that both parties got exposed to both modes several times during the test (10 minutes), but large enough to allow subjects to accommodate to the situation.

The subjects were placed in separate rooms and equipped with head-sets connected to a Voice-over-IP call. On each side, the call is enhanced by the SynFace animated talking head representing the other participant, providing real-time lip-synchronised visual speech animation. The task was to speak about any topic freely for around ten minutes. There were 12 participants making up 6 pairs. None of the participants had any previous knowledge of the experiment setup.

The results were analysed by counting the percentage of times that the turn changed when a speaker paused. The percentage of all utterances followed by a turn change is *larger* under the PASSIVE condition than under the ACTIVE condition for each participant without exception. The difference is significant ($P < .01$), which shows that subjects were consistently affected by the interaction control cues in the talking head. As postinterviews revealed that most subjects never even noticed the gestures consciously, and no subject connected them directly to interaction control, this result shows that it is possible to unobtrusively influence the interaction behaviour of two interlocutors in a given direction—that is to make a person take the floor more or less often—by way of facial gestures in an animated talking head in the role of an avatar.

4. Evaluation Experiments

In the SynFace application, speech intelligibility enhancement is the main function. Speech reading and audio-visual speech intelligibility have been extensively studied by many researchers, for natural speech as well as for visual speech synthesis systems driven by text or phonetically transcribed input. Massaro et al. [7], for example, evaluated visual-only intelligibility of a speaker dependent speech driven

system on isolated words. To date, however, we have not seen any published results on speaker independent speech driven facial animation systems, where the intelligibility enhancement (i.e., audiovisual compared to audio-only condition) has been investigated. Below, we report on two experiments were audiovisual intelligibility of SynFace has been evaluated for different configurations and languages.

The framework adopted in SynFace allows for evaluation of the system at different points in the signal chain shown in Figure 2. We can measure accuracy

(i) at the phonetic level, by measuring the phoneme (viseme) accuracy of the speech recogniser,

(ii) at the face parameter level, by computing the distance between the face parameters generated by the system and the optimal trajectories, for example, trajectories obtained from phonetically annotated speech,

(iii) at the intelligibility level, by performing listening tests with hearing impaired subjects, or with normal hearing subjects and a degraded acoustic signal.

The advantage of the first two methods is simplicity. The computations can be performed automatically, if we assume that a good reference is available (phonetically annotated speech). The third method, however, is the most reliable because it tests the effects of the system as a whole.

Evaluating the Phoneme Recogniser. Measuring the performance at the phonetic level can be done in at least two ways: By measuring the percentage of frames that are correctly classified, or by computing the Levenshtein (edit) distance Levenshtein [34] between the string of phonemes output by the recogniser and the reference transcription. The first method does not explicitly consider the stability of the results in time and, therefore, may overestimate the performance of a recogniser that produces many short insertions. These insertions, however, do not necessarily result in a degradation of the face parameter trajectories, because the articulatory model, the face parameter generation is based on, often acts as a low-pass filter. On the other hand, the Levenshtein distance does not consider the time alignment of the two sequences, and may result in misleading evaluation in the case that two phonetic subsequences that are not co-occurring in time are aligned by mistake. To make the latter measure homogeneous with the correct frames %, we express it in terms of accuracy, defined as $(1 - l/n) \times 100$, where l is Levenshtein (Edit) distance and n the length of the reference transcription.

Intelligibility Tests. Evaluating the intelligibility is performed by listening tests with a number of hearing impaired or normal hearing subjects. Using normal hearing subject and distorting the audio signal has been shown to be a viable simulation of perception by hearing impaired [35, 36]. The speech material is presented to the subjects in different conditions. These may include *audio alone*, *audio and natural face*, *audio and synthetic face*. In the last case, the synthetic face may be driven by different methods (e.g., different versions of the PR that we want to compare). It may also be driven by carefully obtained annotations of the speech material, if the aim is to test the effects of the visual synthesis models alone.

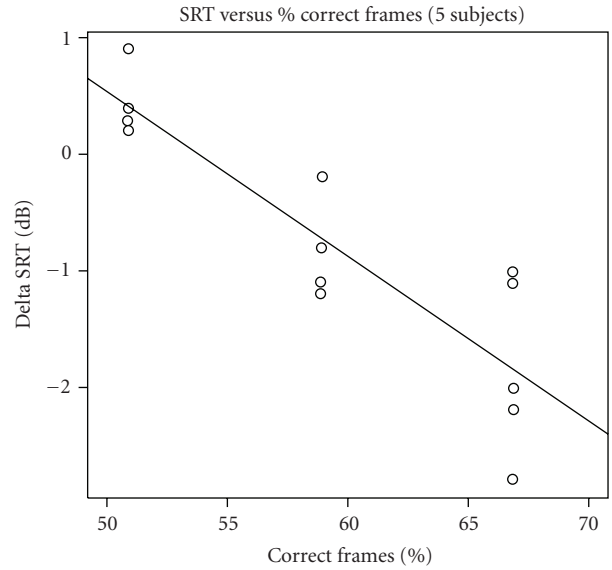

Figure 3: Delta SRT versus correct frames % for three different recognisers (correlation $r = -0.89$) on a 5-subject listening test.

Two listening test methods have been used in the current experiments. The first method is based on a set of carefully designed short sentences containing a number of key-words. The subject's task is to repeat the sentences, and intelligibility is measured in terms of correctly recognised key-words. In case of normal hearing subjects, the acoustic signal may be degraded by noise in order to simulate hearing impairment. In the following, we will refer to this methodology as "key-word" test.

The second methodology Hagerman and Kinnefors [37] relies on the adaptive use of noise to assess the level of intelligibility. Lists of 5 words are presented to the subjects in varying noise conditions. The signal-to-noise ratio (SNR dB) is adjusted during the test until the subject is able to correctly report about 50% of the words. This level of noise is referred to as the Speech Reception Threshold (SRT dB) and indicates the amount of noise the subject is able to tolerate before the intelligibility drops below 50%. Lower values of SRT correspond to better performance (the intelligibility is more robust to noise). We will refer to this methodology as "SRT" test.

SRT Versus Correct Frames %. Figure 3 relates the change of SRT level between audio-alone and SynFace conditions (Delta SRT) to the correct frames % of the corresponding phoneme recogniser. Although the data is based on a small listening experiment (5 subjects), the high correlation shown in the figure motivates the use of the correct frames % measure for developmental purposes. We believe, however, that reliable evaluations should always include listening tests. In the following, we report results on the recent developments of SynFace using both listening tests and PR evaluation.

Table 2: Number of connections in the RNN and correct frames % of the SynFace RNN phonetic classifiers.

Language	Connections	Correct frames %
English	184,848	46.1
Flemish	186,853	51.0
German	541,430	61.0
Swedish	541,250	54.2

These include the newly developed German recogniser, wide- versus narrow-band speech recognition experiments and cross-language tests. All experiments are performed with the real-time, low-latency implementation of the system, that is, the phoneme recogniser uses 30 milliseconds look-ahead length, and the total delay of the system in the intelligibility tests is 200 milliseconds.

4.1. SynFace in German. To extend SynFace to German, a new recogniser was trained on the SpeechDat German recordings. These consist of around 200 hours of telephone speech spoken by 4000 speakers. As for the previous languages, the HTK-based RefRec recogniser Lindberg et al. [38] was trained and used to derive phonetic transcriptions of the corpus. Whereas the recogniser for Swedish, English and Flemish, was trained exclusively on the phonetically rich sentences, the full training set, also containing isolated words, digits, and spellings, was used to train the German models. Table 2 shows the results in terms of correct frames % for the different languages. Note however that these results are not directly comparable because they are obtained on different test sets.

The same synthesis rules used for Swedish are applied to the German system, simply by mapping the phoneme (viseme) inventory of the two languages.

To evaluate the German version of the SynFace system, a small "key-word" intelligibility test was performed. A set of twenty short (4–6 words) sentences from the Göttinger satsztest set [39], spoken by a male native German speaker, were presented to a group of six normal hearing German listeners. The audio presented to the subjects was degraded in order to avoid ceiling effects, using a 3-channel noise excited vocoder shannon et al. [40]. This type of signal degradation has been used in the previous audio-visual intelligibility experiments Siciliano et al. [41] and can be viewed as a way of simulating the information reduction experienced by cochlear implant patients. Clean speech was used to drive SynFace. 10 sentences were presented with audio-only and 10 sentences were presented with SynFace support. Subjects were presented with four training sentences before the test started. The listeners were instructed to watch the screen and write down what they perceived.

Figure 4 summarises the results for each subject. The mean score (% correctly recognised key-words) for the audio only condition was extremely low (2.5%). With SynFace support, a mean score of 16.7% was obtained. While there was a large intersubject variability, subjects consistently showed to take advantage of the use of SynFace. An ANOVA analysis shows significant differences ($P < .01$) between the audio-alone and SynFace conditions.

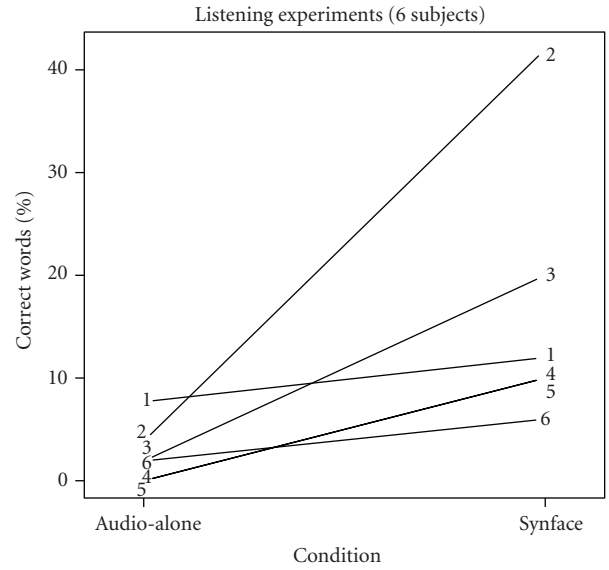

Figure 4: Subjective evaluation results for German version of SynFace (% correct word recognition).

4.2. Narrow- Versus Wide-Band PR. In the Hearing at Home 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 was trained on wide-band speech data from the SpeeCon corpus [42]. SpeeCon contains recordings in several languages and conditions. Only recordings in office settings of Swedish were chosen. The corpus contains word level transcriptions, and annotations for speaker noise, background noise, and filled pauses. As in the SpeechDat training, the silence at the boundaries of every utterance was reduced, in order to improve balance between the number of frames for the silence class and for any other phonetic class. Differently from the SpeechDat training, NALIGN Sjolander [19] was used in order to create time aligned phonetic transcriptions of the corpus based on the orthographic transcriptions.

The bank of filters, used to compute the MFCCs that are input to the recogniser, was defined in a way that the filters between 0 and 4 kHz coincide with the narrow-band filterbank definition. Additional filters are added for the upper 4 kHz frequencies offered by the wide-band signal.

Table 3 shows the results for the network trained on the SpeeCon database. The results obtained on the SpeechDat material are also given for comparison. Note, however, that these results cannot be compared directly because the tests were performed on different test sets.

Table 3: Comparison between the SpeechDat telephone quality (TF), SpeeCon narrow-band (NB) and SpeeCon wide-band (WB) recognisers. Results are given in terms of correct frames % for phonemes (ph) and visemes (vi), and accuracy.

Database	SpeechDat	SpeeCon	
Data size (ca. hours)	200	40	
Speakers (#)	5000	550	
Speech quality	TF	NB	WB
Sampling (kHz)	8	8	16
Correct frames (% ph)	54.2	65.2	68.7
Correct frames (% vi)	59.3	69.0	74.5
Accuracy (% ph)	56.5	62.2	63.2

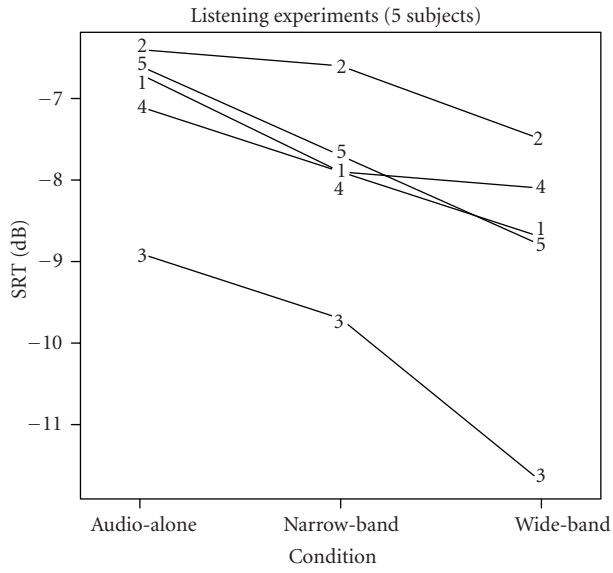

Figure 5: Speech Reception Threshold (SRT dB) for different conditions and subjects.

In order to have a more controlled comparison between the narrow- and the wide-band networks for Swedish, a network was trained on a downsampled (8 kHz) version of the same SpeeCon database. The middle column of Table 3 shows the results for the networks trained and tested on the narrow-band (downsampled) version of the SpeeCon database. Results are shown in terms of % of correct frames for phonemes, visemes, and phoneme accuracy.

Finally, a small-scale "SRT" intelligibility experiment, was performed in order to confirm the improvement in performance that we see in the wide-band case. 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 5 shows the SRT levels obtained in the different conditions, where each line corresponds to a subject. An ANOVA analysis and successive multiple comparison analysis confirm that there is a significant decrease (improvement) of SRT ($P < .001$) for the wide-band recogniser over the narrow-band trained network and the audio-alone condition.

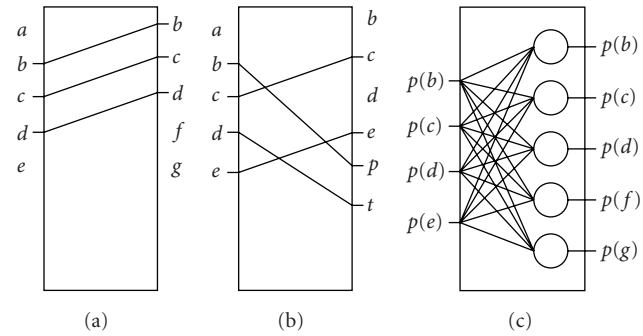

Figure 6: Structure of the language models mappers. (a) Phonetic mapper: only identical phonemes are matched between the two languages, (b) best match mapper: the phoneme are blindly matched in a way that recognition results are maximised, (c) linear regression mapper: posterior probabilities for the target language are estimated from the outputs of the phonetic recogniser.

Table 4: correct frames % for different languages (columns) recognised by different models (rows). The languages are: German (de), English (en), Flemish (fl) and Swedish (sv). Numbers in parentheses are the % of correct frames for perfect recognition, given the mismatch in phonetic inventory across languages.

	de	sv	fl	en
de	61.0 (100)	30.3 (82.6)	27.1 (73.5)	26.2 (71.6)
sv	31.5 (86.2)	54.2 (100)	26.3 (72.1)	23.5 (74.2)
fl	34.2 (85.7)	31.6 (77.9)	51.0 (100)	26.9 (69.8)
en	24.5 (74.6)	23.7 (72.3)	21.5 (66.8)	46.1 (100)

4.3. Multilinguality. 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 illustrated in Figure 6. In the first case (Figure 6(a)) we relay 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.

Table 4 presents the correct frames % when only matching phonetic symbols are considered. The numbers in parentheses show the highest possible performance, given the fact that some of the phonemes in the test set do not exist in the recognition model. As expected, the accuracy of recognition drops drastically when we use cross-language models. This could be considered as a lower bound to performance.

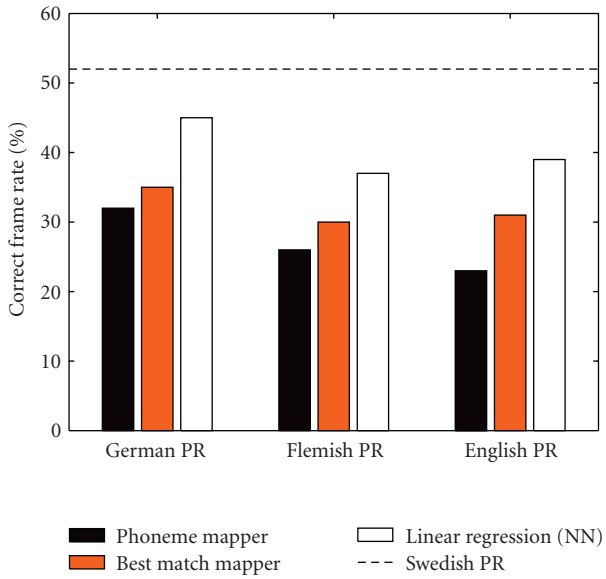

Figure 7: correct frames % for the three mappers when the target language is Swedish. The dashed line represents the results obtained with the Swedish phoneme recogniser (language matching case).

The second mapping criterion depicted in Figure 6(b), considers as correct the association between model and target phonemes that was most frequently adopted by the recognition models on that particular target language. If we consider all possible maps between the set of model phonemes and the set of target phonemes, this corresponds to an upper bound of the results. Compared to the results in Table 4, this evaluation method gives about 10% increased correct frames in average. In this case, there is no guarantee that the chosen mapping bares phonetic significance.

The previous mappings were introduced to allow for simple cross-language evaluation of the phonetic recognisers. Figure 6(c) shows a mapping method that is more realistic in terms of system performance. In this case we do not map between two discrete sets of phonemes, but, rather, between the posterior probabilities of the first set and the second set. 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. Figure 7 shows the results when the target language is Swedish, and the phoneme recognition (PR) models were trained on German, English and Flemish. Only 30 minutes of an independent training set were used to train the linear regression mapper. The performance of this mapper is above the best match results, and comes close to the Swedish PR results (dashed line in the figure) for the German PR models.

5. Conclusions

The purpose of SynFace is to enhance spoken communication for the hearing impaired, rather than solving the acoustic-to-visual speech mapping per se. The methods employed here are, therefore, tailored to achieving this goal in the most effective way. Beskow [18] showed that, whereas data-driven visual synthesis resulted in more realistic lip movements, the rule-based system enhanced the intelligibility. Similarly, mapping from the acoustic speech directly into visual parameters is an appealing research problem. However, when the ambition is to develop a tool that can be applied in real-life conditions, it is necessary to constrain the problem. The system discussed in this paper

(i) works in real time and with low latency, allowing realistic conditions for a natural spoken communication,

(ii) is light-weight and can be run on standard commercially available hardware,

(iii) is speaker independent, allowing the user to communicate with any person,

(iv) is being developed for different languages (currently, Swedish, English, Flemish, and German are available),

(v) is optimised for different acoustic conditions, ranging from telephone speech quality to wide-band speech available in, for example, Internet communications and radio/TV broadcasting,

(vi) is being extensively evaluated in realistic settings, with hearing impaired subjects or by simulating hearing impairment.

Even though speech intelligibility is the focus of the SynFace system, extra-linguistic aspects of speech communication have also been described in the paper. Modelling nonverbal gestures proved to be a viable way of enhancing the turn-taking mechanism in telephone communication.

Future work will be aimed at increasing the generality of the methods, for example, by studying ways to achieve language independence or by simplifying the process of optimising the system to a new language, based on the preliminary results shown in this paper. Reliably extracting extra-linguistic information, as well as synthesis and evaluation of nonverbal gestures will also be the focus of future work.

Acknowledgments

The work presented here was funded in part by European Commission Project IST-045089 (Hearing at Home) and Swedish Research Council Project 621-2005-3488 (Modelling multimodal communicative signal and expressive speech for embodied conversational agents).

References

[1] M. Tamura, T. Masuko, T. Kobayashi, and K. Tokuda, "Visual speech synthesis based on parameter generation from hmm: speech-driven and text-and-speechdriven approaches," in *Proceedings of the Auditory-Visual Speech Processing (AVSP '98)*, pp. 221–226, 1998.

[2] S. Nakamura and E. Yamamoto, "Speech-to-lip movement synthesis by maximizing audio-visual joint probability based on the EM algorithm," *Journal of VLSI Signal Processing*

Systems for Signal, Image, and Video Technology, vol. 27, no. 1-2, pp. 119–126, 2001.

[3] Z. Wen, P. Hong, and T. Huang, "Real time speech driven facial animation using formant analysis," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '01)*, pp. 817–820, 2001.

[4] S. Al Moubayed, M. De Smet, and H. Van Hamme, "Lip synchronization:from phone lattice to PCA eigenprojections using neural networks," in *Proceedings of the Biennial Conference of the International Speech Communication Association (Interspeech '08)*, Brisbane, Australia, 2008.

[5] G. Hofer, J. Yamagishi, and H. Shimodaira, "Speech-driven lip motion generation with a trajectory hmm," in *Proceedings of the Biennial Conference of the International Speech Communication Association (Interspeech '08)*, Brisbane, Australia, 2008.

[6] T. Ohman and G. Salvi, "Using HMMs and ANNs for mapping acoustic to visual speech," *TMH-QPSR*, vol. 40, no. 1-2, pp. 45–50, 1999.

[7] D. Massaro, J. Beskow, M. Cohen, C. Fry, and T. Rodgriguez, "Picture my voice: audio to visual speech synthesis using artificial neural networks," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (ISCA '99)*, 1999.

[8] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," in *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 388–398, ACM, New York, NY, USA, 2002.

[9] K. Liu and J. Ostermann, "Realistic facial animation system for interactive services," in *Proceedings of the Biennial Conference of the International Speech Communication Association (Interspeech '08)*, Brisbane, Australia, 2008.

[10] M. Cohen and D. Massaro, "Models and techniques in computer animation," in *Modeling Coarticulation in Synthetic Visual Speech*, vol. 92, Springer, Tokyo, Japan, 1993.

[11] M. Železný, Z. Krňoul, P. Císař, and J. Matoušek, "Design, implementation and evaluation of the Czech realistic audio-visual speech synthesis," *Signal Processing*, vol. 86, no. 12, pp. 3657–3673, 2006.

[12] B. J. Theobald, J. A. Bangham, I. A. Matthews, and G. G. Cawley, "Near-videorealistic synthetic talking faces: implementation and evaluation," *Speech Communication*, vol. 44, no. 1–4, pp. 127–140, 2004.

[13] N. Kitawaki and K. Itoh, "Pure delay effects on speech quality in telecommunications," *IEEE Journal on Selected Areas in Communications*, vol. 9, no. 4, pp. 586–593, 1991.

[14] M. McGrath and Q. Summerfield, "Intermodal timing relations and audio-visual speech recognition by normal-hearing adults," *Journal of the Acoustical Society of America*, vol. 77, no. 2, pp. 678–685, 1985.

[15] F. I. Parke, "Parameterized models for facial animation," *IEEE Computer Graphics and Applications*, vol. 2, no. 9, pp. 61–68, 1982.

[16] J. Beskow, "Rule-based visual speech synthesis," in *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech '95)*, pp. 299–302, Madrid, Spain, 1995.

[17] T. Gjermani, *Integration of an animated talking face model in a portable device for multimodal speech synthesis*, M.S. thesis, Department for Speech, Music and Hearing, KTH, School of Computer Science and Communication, Stockholm, Sweden, 2008.

[18] J. Beskow, "Trainable articulatory control models for visual speech synthesis," *International Journal of Speech Technology*, vol. 7, no. 4, pp. 335–349, 2004.

[19] K. Sjolander, "An HMM-based system for automatic segmentation and alignment of speech," in *Proceedings of Fonetik*, pp. 93–96, Umeå, Sweden, 2003.

[20] G. Salvi, "Truncation error and dynamics in very low latency phonetic recognition," in *Proceedings of the Non Linear Speech Processing (NOLISP '03)*, Le Croisic, France, 2003.

[21] A. J. Robinson, "Application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 298–305, 1994.

[22] P. J. Werbos, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[23] H. Bourlard and N. Morgan, "Continuous speech recognition by connectionist statistical methods," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 893–909, 1993.

[24] N. Strom, "Development of a recurrent time-delay neural net speech recognition system," *TMH-QPSR*, vol. 26, no. 4, pp. 1–15, 1992.

[25] K. Elenius, "Experiences from collecting two Swedish telephone speech databases," *International Journal of Speech Technology*, vol. 3, no. 2, pp. 119–127, 2000.

[26] G. Salvi, "Dynamic behaviour of connectionist speech recognition with strong latency constraints," *Speech Communication*, vol. 48, no. 7, pp. 802–818, 2006.

[27] K. G. Munhall, J. A. Jones, D. E. Callan, T. Kuratate, and E. Vatikiotis-Bateson, "Visual prosody and speech intelligibility: head movement improves auditory speech perception," *Psychological Science*, vol. 15, no. 2, pp. 133–137, 2004.

[28] Y. Cao, W. C. Tien, P. Faloutsos, and F. Pighin, "Expressive speech-driven facial animation," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1283–1302, 2005.

[29] U. Hadar, T. J. Steiner, E. C. Grant, and F. C. Rose, "Kinematics of head movements accompanying speech during conversation," *Human Movement Science*, vol. 2, no. 1-2, pp. 35–46, 1983.

[30] J. Beskow, B. Granström, and D. House, "Analysis and synthesis of multimodal verbal and non-verbal interaction for animated interface agents," in *Proceedings of the International Workshop on Verbal and Nonverbal Communication Behaviours*, vol. 4775 of *Lecture Notes in Computer Science*, pp. 250–263, 2007.

[31] J. Edlund and J. Beskow, "Pushy versus meek-using avatars to in uence turn-taking behaviour," in *Proceedings of the Biennial Conference of the International Speech Communication Association (Interspeech '07)*, Antwerp, Belgium, 2007.

[32] A. Kendon, "Some functions of gaze-direction in social interaction," *Acta Psychologica*, vol. 26, no. 1, pp. 22–63, 1967.

[33] V. Hugot, *Eye gaze analysis in human-human commuication*, M.S. thesis, Department for Speech, Music and Hearing, KTH, School of Computer Science and Communication, Stockholm, Sweden, 2007.

[34] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.

[35] L. E. Humes, B. Espinoza-Varas, and C. S. Watson, "Modeling sensorineural hearing loss. I. Model and retrospective evaluation," *Journal of the Acoustical Society of America*, vol. 83, no. 1, pp. 188–202, 1988.

[36] L. E. Humes and W. Jesteadt, "Models of the effects of threshold on loudness growth and summation," *Journal of the Acoustical Society of America*, vol. 90, no. 4, pp. 1933–1943, 1991.

[37] B. Hagerman and C. Kinnefors, "Efficient adaptive methods for measuring speech reception threshold in quiet and in noise," *Scandinavian Audiology*, vol. 24, no. 1, pp. 71–77, 1995.

[38] B. Lindberg, F. T. Johansen, N. Warakagoda, et al., "A noise robust multilingual reference recogniser based on SpeechDat(II)," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP '00)*, 2000.

[39] M. Wesselkamp, *Messung und modellierung der verstandlichkeit von sprache*, Ph.D. thesis, Universitat Gottingen, 1994.

[40] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[41] C. Siciliano, G. Williams, J. Beskow, and A. Faulkner, "Evaluation of a multilingual synthetic talking face as a communication aid for the hearing impaired," in *Proceedings of the 15th International Conference of Phonetic Sciences (ICPhS '03)*, pp. 131–134, Barcelona, Spain, 2003.

[42] D. Iskra, B. Grosskopf, K. Marasek, H. V. D. Heuvel, F. Diehl, and A. Kiessling, "Speecon—speech databases for consumer devices: database specification and validation," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC '02)*, pp. 329–333, 2002.