Avatar Therapy: an audio-visual dialogue system for treating auditory hallucinations

Mark Huckvale[1], *Julian Leff*[2], *Geoff Williams*[1],

[1] Department of Speech, Hearing and Phonetics, University College London, UK
[2]Department of Mental Health Sciences, University College London, UK

m.huckvale@ucl.ac.uk, j.leff@ucl.ac.uk, geoffrey.williams@ucl.ac.uk

10.21437/Interspeech.2013-107

Abstract

This paper presents a radical new therapy for persecutory auditory hallucinations ("voices") which are most commonly found in serious mental illnesses such as schizophrenia. In around 30% of patients these symptoms are not alleviated by anti-psychotic medication. This work is designed to tackle the problem created by the inaccessibility of the patients' experience of voices to the clinician. Patients are invited to create an external representation of their dominant voice hallucination using computer speech and animation technology. Customised graphics software is used to create an avatar that gives a face to the voice, while voice morphing software realises it in audio, in real time. The therapist then conducts a dialogue between the avatar and the patient, with a view to gradually bringing the avatar, and ultimately the hallucinatory voice, under the patient's control. Results of a pilot study reported elsewhere indicate that the approach has potential for dramatic improvements in patient control of the voices after a series of only six short sessions. The focus of this paper is on the audio-visual speech technology which delivers the central aspects of the therapy.

Index Terms: voice conversion, facial animation, audio-visual speech

1. Introduction

The phenomenon of auditory hallucinations ("hearing voices") is an enduring problem in the treatment of serious mental illness such as schizophrenia. About 30% of people with this diagnosis continue to experience hallucinations and delusions despite treatment with antipsychotic medication [1]. Hearing voices is not only distressing to the sufferers, it also has a serious impact on their carers and members of the public with whom they come into contact. Auditory hallucinations manifest in a number of ways, including voices that speak aloud what the patient is thinking; voices giving a running commentary on the patient's actions or external imagined events; two or more persons conversing, often referring to the patient in the third person; and commands ordering the patient to perform certain actions (often violent). Persistent voices severely limit the patients' ability to concentrate on tasks, and hence hinder attempts at rehabilitation. The direct treatment costs in the United Kingdom are estimated at £2billion annually, while the indirect costs, including loss of employment for the patients and carers, amount to another £2billion [2].

In the past 15 years or so in Britain a number of randomised controlled trials (RCTs) have been conducted to test the value of cognitive-behavioural therapy (CBT) for persistent medication-resistant symptoms of psychosis [3], [4],[5],[6]. While these have shown some effect in reducing auditory hallucinations, they have been criticised on grounds of experimental design. One more recent RCT of CBT, while not affecting the frequency or intensity of auditory hallucinations, did succeed in reducing the power of the dominant voice as perceived by the patients, and their distress [7].

When asked about the worst aspect of hearing persecutory voices, many patients report 'the feeling of helplessness' it induces. Those who are able to establish a dialogue with their voice feel much more in control and their suffering is consequently reduced. Many voice hearers also visualise a face associated with their voice. This can be the face of someone known to them, a well-known personality or an imaginary figure, or perhaps representing an angel, devil or other religious or mythical figure. One means by which a patient could be helped to gain control of their voice is by creating a virtual avatar that represents the person they believe talks to them, and then allowing the avatar to progressively come under the patient's control. To embody the voice within the context of an avatar therefore, is a natural step to make from a clinical point of view.

Virtual reality (VR) techniques have previously been explored for modelling the psychotic episodes involved in schizophrenia. Banks *et al*[8] report a sophisticated VR environment for simulating auditory and visual hallucinations and motivate its use in medical education, but have not tested their system in a clinical setting, nor attempted to individualise features of the voices.

This paper describes a highly novel speech-technology system for delivering CBT-based therapy for auditory hallucinations, which we term Avatar Therapy. The basis of the system is a series of components for creating and operating an individualised avatar that "speaks" in the voice that the patient hears, and visually resembles the face that the patient perceives. (In cases where the patient does not clearly perceive a face, he/she is asked to choose a face which they would feel comfortable talking to). Characteristics of the avatar are chosen in consultation with each patient beforehand using a series of computerised tools. A sample set of avatar images chosen by patients is shown in Figure 1.

Figure 1: Samples of actual avatar characters

During the therapy sessions, the therapist and the patient sit in front of a computer screen in separate rooms and communicate via a two-way (half-duplex) audio link. The avatar's utterances are voiced by the therapist, and the patient's responses to the avatar are fed back (audio only), so that the patient can interact with and challenge the avatar. The voice of the avatar is produced by modifying the therapist's speech in real time, so that the therapist's speech is voiced by

Copyright © 2013 ISCA

the avatar in the simulated voice on the client's computer with lip synchronisation. Over the same channel, the therapist can also communicate instructions or advice and encouragement to the patient in his/her own voice, as would be the case in a standard therapy session. Results of a small-scale RCT study funded by NIHR are reported in [9] and further discussed in [10].

2. System design considerations

2.1. Generating the avatar utterances

An important parameter in the design of a dialogue system with a synthetic voice is the method of generating the synthetic utterances: either directly synthesising speech with the desired characteristics, or by transforming natural speech to produce a different quality to that of the original speaker. Since the content of the therapist's utterances to be voiced by the avatar cannot be known in advance, a standard TTS system was of no use here. Typing in text during the therapy sessions would introduce frequent unacceptably long delays, disrupting the flow of the dialogue. Therefore, to allow as near real-time interaction as possible, a system based on voice conversion rather than speech synthesis was selected.

The design parameters of the experimental study required that the target voice and face be generated in an initial enrolment session with each patient, lasting no more than an hour. Ideally this would be followed by the first of the series of therapy sessions, meaning that the avatar must be obtained in its final form during the enrolment session. Subjects were recruited continuously throughout the study and so the target voices and voices were not available in advance of the technology development.

In conventional applications of voice conversion where the set of target speakers is known in advance, a training procedure is applied to produce the mapping, or transform, from the source to the target voice. This requires the availability of speech training data from both source and target speakers. In our case then, the major technical problem was the fact that the target voices are not known until the patients have been enrolled and, of course, no actual samples could be obtained in any case. This means that the mappings cannot be trained individually and that a set of predetermined voice transforms must be developed instead, using the therapist's voice as the source. While this limits the range of voices that can be delivered in good quality, a variety of different voices can still be produced, provided some additional means of fine-tuning the transforms is available..

2.2. System components for enrolment and therapy

The system therefore comprises a combination of off-line and on-line procedures and includes the following components:

- Real-time voice conversion system
- Voice and face customization (enrolment) systems
- Customizable facial animation system with real-time lip-synching
- Two-way audio channel with switching between therapist and avatar voice

In the off-line, or enrolment procedures, the patient chooses the appropriate voice and face using a set of computer-based tools, assisted by a trained operator, supervised by the therapist. Some of these tools have been developed by the authors, while others are commercially available products or software toolkits that we have customised to suit the needs of the task.

3. Real-time voice conversion system

3.1. Technology

The voice conversion technology was broadly based on the approach described by Stylianou [11]. In this approach, speech is decomposed into two elements: a time-varying filter and a residual excitation waveform. Spectral transformations of the filter combined with pitch modification of the residual allow the synthesis of versions of an utterance in which the characteristics of the speaker seem to have changed.

Decomposition of the speech is performed by linear prediction, with the prediction filter being estimated over short windowed sections of the signal 30ms in duration, and overlapping by 10ms. The prediction error of the filter is then calculated separately for each window, and then overlapped and added to create the excitation residual waveform. To make the prediction coefficients more amenable to spectral mapping, the prediction coefficients are Fourier transformed into a 256 point amplitude response. This transformation is just a mathematical convenience, allowing the filter response to be adjusted by a spectral warping function for transformation of the voice characteristics. For synthesis, the warped spectrum is then converted back to predictor coefficients. With a uniform spectral warping, the conversion to/from the amplitude response does not introduce noticeable signal distortion.

Spectral manipulation of the filter response is performed by a set of linear transforms. In this work, a set of 8 transforms is used for each target voice, the exact forms of which are found during the training and customization procedure described below. To select how the transforms are used for a given stretch of speech signal, the signal spectral envelope for the source speaker is modeled using a Gaussian Mixture Model (GMM) of 8 mixtures. Performance can be improved with 64 mixtures ([11],[12]), but as our voice conversion system is designed for flexibility and real-time operation, with no concept of "accurately" recreating a specific voice, 8 mixtures constitute the best compromise between the required quality and minimal computational load. Each mixture is then associated with a linear transform and the final transform applied to the filter is found from the sum of transforms weighted according to the mixture probability for the corresponding source speech signal envelope. See [11] for a mathematical description.

For pitch scaling, the residual signal from the source speaker is first resampled to change the fundamental frequency by a constant factor. To correct for the change in duration caused by the resampling, the residual is then compressed or stretched in time to restore the original duration using Waveform-Similarity Overlap-Add (WSOLA, see [13]).

3.2. Training

The input to the voice customization process is a large number of trained voice transformations generated between the single source speaker (the therapist) and a number of training speakers. For this work we used 55 speakers taken from the Accents of British English corpus [14] and 27 speakers taken from the UCL Speaker Variability corpus [15]. In total there were 40 male and 42 female target speakers. In each case we used a selection of 20 sentences spoken by each speaker and a matching set of 20 sentences spoken by the source speaker.

The spoken materials from the source speaker were used to build an 8-mixture GMM based on MFCC features of the signal [16]. This GMM was then used to train all 82 linear transform sets using the Stylianou method. This first involved, for the source speaker and each target speaker, a temporal alignment between each matching pair of sentences. The MFCC vectors were used for this alignment in combination with a dynamic programming algorithm. Then for each matched pair of signal sections, the predictor coefficients and LPC spectrum were calculated and the optimum frequency mapping found for that pair of frames using a dynamic programming algorithm. Given all the mappings for all paired frames in all sentences together with the set of GMM mixture probabilities for each source frame, it is then possible to find the set of 8 average transforms which minimize the mean squared transformation error. In addition, the mean fundamental frequency of each speaker and the source speaker were measured. Thus for each target speaker, the training procedure generates a set of 8 linear frequency transforms and a pitch scaling factor, which taken together makes the best attempt at morphing the source speaker's voice characteristics to the target speaker.

4. Enrolment procedures

4.1. Face selection and customization

The face selection procedure is based around the FaceGen© Modeller software developed by Singular Inversions. This allows a virtually infinite range of 3-D faces to be created and saved in various file formats. As a starting point in the selection, a visual array or "palette" of around 250 distinct faces was created, covering a broad range of features. The faces are chosen to cover a broad range of face types which vary in gender, age and ethnic group , as well as various common hairstyles (Figure 2).

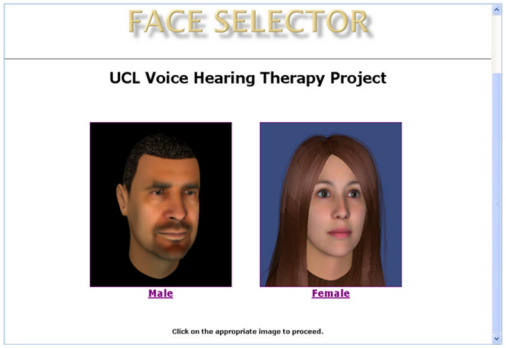

Figure 2: Front page of on-line Face Selector

A limited range of hairstyles and accessories is available. Each of these files was then saved both as a FaceGen model file and as a JPEG image file. The image files were compiled into web pages which were linked together to form a web site for browsing the initial palette of faces.

Each of the images on the face selector site has an associated FaceGen model file, stored in a parallel directory structure. Once the client has chosen a suitable starting face, the corresponding face model file is easily located for further refined in FaceGen as the client wishes. The resulting face image is exported as a set of animation targets in the .OBJ 3D graphics format. These are essentially a set of image and texture files for the pre-selected set of visemes, which are used in the facial animation system (see Section 5 below).

4.2. Voice enrolment

In the pilot study customization of the voice was performed in two stages, but in the present version the process is integrated into a single application. First a sample of about 20 voice transforms for the appropriate gender were chosen from the training set, and the client was asked to select which of these had the closest speaker characteristics to the required voice. In the selection screen, each speech bubble represents a different voice and when clicked, it plays out an audio file with lip-sync in the voice which it represents (see Figure 3).

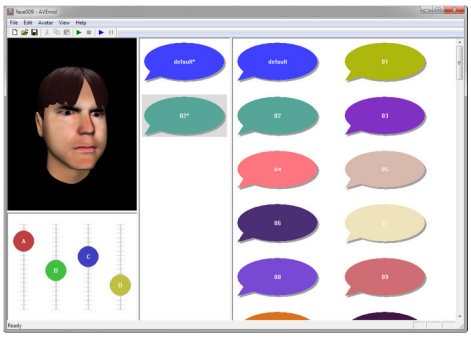

Figure 3: Voice enrolment interface (current)

In a second stage, the client manipulated the chosen transform set using a small set of sliders which altered the pitch scaling factor and the transform coefficients (now integrated into the screen of Figure 3). To allow for manipulation of the transforms using a small number of sliders, principal components analysis was performed on the 82 transforms estimated during training. This reduced a chosen transform vector to a mean transform vector plus a number of orthogonal change vectors. Each slider manipulates the amount by which each change vector was added in to the mean vector. A separate slider manipulates the pitch scaling factor. The user interface was designed to allow for the client to experiment with multiple variants of the chosen voice, and to keep copies of promising alternatives, such that at the end of the process, the best transform possible could be saved.

4.3. Performance issues

In the development phase, the voice and face selection procedures were trialed with three voice hearers who did not take part in the subsequent study. They reported a likeness accuracy of around 60-70% with respect to the target face. A similar figure was found with the patients who took part in the study. In a few cases however, the effect was more dramatic. A small number of patients, when shown the avatar for the first time, actually found the realism so great as to be unable to view it for very long.

Given the technical limitations and the requirement for all the voices to be derivable from a single source (a male speaker around 70 years of age with a bass voice), it proved difficult in some cases to match the voice closely. Unsurprisingly, the greatest difficulty was found in matching female voices. Overall, patients reported the accuracy of the match to be in the range 60-90%. The duration of the process depended on such factors as the communicative ability of the client, the level of desired detail and the degree of customization required. It typically takes about 20 minutes, rarely less than 15 or more than 30 minutes.

4.4. Delivery system

Delivery of the real-time converted voice was performed using custom software running on two PCs connected over a network: a "server" process running at the clinician end, and a "client" process running at the patient end. A schematic illustration of the experimental layout is shown in Figure 4 (supplied separately for reasons of space).

The server process captures audio from the therapist's microphone, and performs all the signal processing and speech conversion. The output speech is then transferred over the network to the client computer. The client process replays the converted speech and captures the patient's speech, uploading it to the therapist's computer. The therapist uses two "push-to-talk" buttons which control whether his/her natural voice or the converted voice is output to the client. The therapist listens on headphones, so that the client's responses are not fed back via the therapist's own microphone. Similarly, the client's microphone is muted while audio is being sent from the server, again to avoid feedback through the system. This has a limited but acceptable effect on the dialogue.

Crucial to creating the separation between the therapist and the avatar is ensuring that only the avatar's utterances are passed through the talking head. Since both signals are carried over the same channel, this was achieved by capturing speech in stereo at the server (therapist) end and passing the unprocessed and the converted audio through the left and right channels respectively. Only the right channel was then passed to the recognizer and lip-sync engine in the animation client, which achieves the desired effect.

Finally, the client end is equipped with a panic button which turns off the avatar immediately when pressed, displays a scenic image and plays soothing music. Few clients found the need to use this facility but its presence alone proved helpful in allaying some clients' anxiety.

4.5. Discussion

The strengths of the audio system were that voice customization and real-time delivery of a customized voice were achieved. When the target voice was male and the transform not too large, the quality of transformation was good and relatively undistorted. The main weakness was that the range of available voices of good quality was somewhat limited; and when the required transforms were too great, there was also a loss in quality. The female voices generated from a male therapist's voice sounded somewhat strained and unconvincing to some patients.

In addition, because all of the audio processing was performed using standard sound cards and audio drivers, the processing system added about 50ms of delay into the dialogue. Additional, variable delays introduced by the network link caused more serious problems that interfered with the conduct of the session. For this reason, the system has since been re-implemented to work with a direct audio link, with the audio processing re-engineered to use the Windows core audio platform, resulting in much lower latencies.

5. Facial animation platform

5.1. Implementation

The animation application was based on the real-time lip-sync SDK from Annosoft LLC [17]. Typically used for animating 3D talking characters in video games or films, this toolkit allows 3D characters designed in other applications to be imported in a range of formats and passed to a real-time phonetic speech recognizer which outputs lip movements. The system can work from both direct audio input and from pre-recorded speech. Character information and animation targets are loaded from a simple plain-text configuration file linked to a set of graphic files, and can be pre-compiled into a binary format for faster loading and display. In the present application, a set of 12 visemes (see e.g. [18]) was used, including one for silence. This was found to be the best compromise between the accuracy of the mouth movements and the smoothness and speed of the response. The configuration file for each character simply specifies which animation target of the 3D model is to be used for each viseme, and which audio transform to use. Creation of the configuration files is built into the enrolment programs, based on a predetermined list of phoneme-viseme mappings.

Estimation of lip shapes improves if the recognizer has access to longer stretches of speech, which results in a trade-off between lip-synching accuracy and response latency. The best compromise was obtained with a latency value of 150ms.

In practice, limitations were found in the range of effects that were achievable in the original development timescale (7 months) with the combination of the two tools. Some of the hairstyles did not render well in the animation stage and compromises had to be made to find the most acceptable combination of available features. Further development work is ongoing to address these limitations, focusing on suitable graphical tools for modifying the stored faces prior to animation. Of course, allowing such a refinement stage implies sufficient time between enrolment and the initial therapy session, which is ultimately a clinical decision. In any case, the level of accuracy that is needed to create the appropriate level of realism is an open question, and outside the scope of the original study.

6. Conclusions

To our knowledge, no other study exists of the treatment of auditory hallucinations using computer generated avatars. The technology used in the original pilot study was developed in a short space of time, and is now being further developed as part of a 3-year clinical trial involving a larger group of patients (140), including its extension to other languages. All the basic components are well-known in the speech field, but their combination and adaptation to suit the particular application are novel. In practice, it is relatively straightforward and cost-effective to implement in a clinical setting, requiring only two standard desktop computers connected by an audio cable. Patients readily grasp the idea of the system and are motivated by it. From the psychiatric point of view, the introduction of an avatar allows researchers to study the relationship between the patients and their voices, at first hand and for the first time.

Clinical results have so far proved highly promising, with dramatic results in a small number of patients [9], [10]. Funding constraints initially allowed for only a single (male) therapist; the present study includes both male and female, thus expanding the set of available voices. The major aim of the new study is to replicate avatar therapy in a new setting with different therapists, to establish whether therapeutic benefits can be obtained independently by any trained therapist. Other research questions include testing the effect of the closeness of the match in the voice and face, the level of realism required to obtain the best therapeutic effect, whether the creation of the avatar is itself therapeutic, and which sub-groups of patients is most likely to benefit from the therapy.

7. References

[1] Kane, J. M. (1996) Treatment resistant schizophrenic patients. Journal of Clinical Psychology, 57 (suppl. 9), 35-40.

[2] Barbato, A. (1998) Schizophrenia and Public Health. Geneva: World Health Organization.

[3] Tarrier, N., Beckett, R. et al (1993) A trial of two cognitive behavioural methods of treating drug-resistant psychotic symptoms in schizophrenic patients. I: Outcome. British Journal of Psychiatry, 162, 524-532.

[4] Drury, V., Birchwood, M., Cochrane, R. & Macmillan, F. (1996) Cognitive therapy and recovery from acute psychosis: a controlled trial. I. Impact on psychotic symptoms. British Journal of Psychiatry, 169, 593-601.

[5] Kuipers, E., Fowler, D. et al (1998) London-East Anglia randomised controlled trial of cognitive-behavioural therapy for psychosis. III: follow-up and economic evaluation at 18 months. British Journal of Psychiatry, 173, 61-68.

[6] Sensky, T., Turkington, D. et al (2000) A randomised controlled trial of cognitive behavioural therapy for persistent symptoms in schizophrenia resistant to medication. Archives of General Psychiatry, 57, 165-172.

[7] Trower, P., Birchwood, M. et al (2004) Cognitive therapy for command hallucinations: randomised controlled trial. British Journal of Psychiatry, 184, 312-320.

[8] Banks, J., Ericksson, G., Burrage, K., Yellowlees, P., Ivermee, S. &Tichon, J., "Constructing the hallucinations of psychosis in Virtual Reality", Journal of Network and Computer Applications 27 (2004) 1–11.

[9] Leff, J., Williams, G., Huckvale, M., Arbuthnot, M., & Leff, A. P. (n.d.) (2013) Silencing voices: a proof-of-concept study of computer-assisted therapy for medication-resistant auditory hallucinations. British Journal of Psychiatry (in press).

[10] Leff, J., Williams, G., Huckvale, M., Arbuthnot, M., & Leff, A. P. (2013). Avatar therapy for persecutory auditory hallucinations: What is it and how does it work?. *Psychosis: Psychological, Social and Integrative Approaches*. (online, March 2013)

[11] Stylianou, Y., Cappé, O., Moulines, E., "Statistical Methods for Voice Quality Transformation", Proc. EuroSpeech 1995, Madrid, Spain, 1995.

[12] Toda, T., Black, A. W., & Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans., Audio, Speech, and Language Processing, 15(8), 2222-2235.

[13] Verhelst, W, &Roelands, M., "An overlap-add technique based on waveform similarity (WSOLA) for high-quality time-scale modification of speech", IEEE Conference Acoustics, Speech and Signal Processing (1993) 554-557.

[14] D'Arcy, S.M., Russell, M.J., Browning, S.R. and Tomlinson, M.J., "The Accents of the British Isles (ABI) Corpus", Proc. Modélisations pour l'Identification des Langues, MIDL Paris, 115-119, 2004.

[15] Markham, D. & V. Hazan, "The UCL Speaker Database", Speech, Hearing and Language: UCL Work in Progress, vol. 14, p.1-17, 2002.

[16] Davis, S.B., &Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously-spoken sentences", IEEE Trans. Acoustics, Speech and Signal Processing, 28:357-366, 1980.

[17] Annosoft real-time lip-sync SDK (see http://www.annosoft.com/microphone-lipsync).

[18] Massaro, D. W. (1998). Perceiving talking faces: From speech perception to a behavioral principle. Cambridge, Massachusetts: MIT Press.