Getting Closer – Tailored Multi-Modal Human-Computer Interaction

Florian Metze
Deutsche Telekom Laboratories
Berlin; Germany
florian.metze@telekom.de

Roman Englert
Deutsche Telekom Laboratories at Ben-Gurion University of the Negev
Be'er Sheva; Israel
roman.englert@telekom.de

Udo Bub
Deutsche Telekom Laboratories
Berlin; Germany
udo.bub@telekom.de

Felix Burkhardt
T-Systems Enterprise Services GmbH, SSC ENPS
Berlin; Germany
felix.burkhardt@t-systems.com

Bernhard Kaspar
T-Systems Enterprise Services GmbH, SSC ENPS
Darmstadt; Germany
bernhard.kaspar@t-systems.com

Joachim Stegmann
T-Systems Enterprise Services GmbH, SSC ENPS
Darmstadt; Germany
joachim.stegmann@t-systems.com

ABSTRACT
This paper outlines our vision of an advanced multi-modal call center using avatar technology, which adapts content, presentation, and interaction strategy to properties of the caller such as age, gender, and emotional state. User studies on Interactive Voice Response (IVR) systems have shown that these properties could be used effectively to "tailor" services to users who do not maintain personal preferences, e.g. because they do not use the service on a regular basis. To achieve individualization of services, we focus on the analysis of a caller's voice, as it is available in all our scenarios.

In this paper, we present a survey of our current work on component technologies such as emotion detection, age and gender recognition, and real-time avatar animation. We also present results of usability and acceptability tests as well as an architecture to integrate these technologies in a future multi-modal user interface. This will be consistent across several devices and can be used to access either IVR systems or "live" agents.

Author Keywords
non-verbal vocal interaction, augmented user interfaces, adaptive systems, multi-modal dialog systems

ACM Classification Keywords
H.5.1 Multimedia Information Systems

INTRODUCTION
With the increasing availability of interactive TV over broadband IP networks in homes and the advent of mobile TV solutions combined with novel device-specific input methods, it will be possible to transform conventional call centers into multi-modal support centers within the next few years. This is true for call centers involving Human agents as well as services based on automatic interactive voice response (IVR) systems. Entertainment, services, and support can be augmented by using advanced image processing technologies enabling an avatar representation of Humans. These advanced output technologies of course should be matched by appropriate sensors and appropriate processing on the input side, in order to be able to adapt content, presentation, and interaction strategy not only to user preferences or the capabilities of the terminal device, but also to caller properties such as age, gender, and emotional state. Users calling from a mobile phone will therefore be using the adapted dialog with a voice-only interface, while users calling from a video-conferencing capable device (e.g. an IPTV set-top box or a laptop) could make use of all the encanced features of the proposed system. We believe that this approach will greatly enhance the accessibility and acceptance of services by providing assistance particularly for infrequent users or less experienced, inhomogeneous user groups.

In this paper, we present several technologies which we plan to integrate in an animated avatar system, to be steered either by synthetically generated movements or by live body motion of a real operator in a call center. While the final target of this research is a fully automated "agent" for (among others) customer support applications, hybrid applications, in which at least one participant of a video conference is "hidden" behind an avatar, are also possible: avatars steered by Human operators instead of algorithms can be used to protect privacy and can even be used for fun. Also, it is possible to imagine situations in which a Human might seamlessly assume control of an avatar previously controled algorithmically. Our experiments are conducted on platforms very close to production IVR systems to facilitate the migration of existing portals and also to allow for the deployment of individual components in isolation.

This paper surveys results on the usability and acceptability of avatars and dialog systems tailored for specific age and gender groups in an audio-only setting. We therefore plan to integrate in these enriched call centers software for automatic speech recognition (ASR, potentially on both caller and agent side) and modules for the detection of age, gender, and emotion of the caller. As we consider speech our

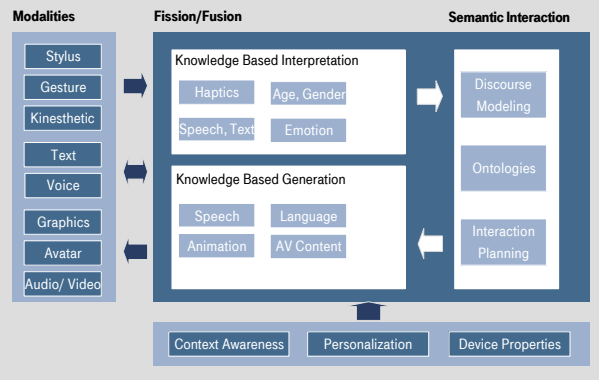

Figure 1. The planned overall system architecture: in this paper, we are specifically concerned with the "age, gender and emotion" interpretation modules using the "voice" modality and the effects of their output on "semantic interaction" (or speech dialog).

"main" modality, analysis will be based on the speech signal alone. Knowledge about age and gender allows to generate user-group specific variants of our telephony-based dialog systems, which could be deployed to avoid having to deal with "angry" customers in the first case. Our experience shows that use patterns (e.g. number and type of words used) of IVR systems vary noticeably for different age/ gender groups [4], so that ASR grammars can be specialized to the user group detected.

SYSTEM ARCHITECTURE

The individual components of the system described in this paper have already been developed and optimized individually. This section describes the avatar animation system and the architecture of our speech dialog system, which will eventually evolve into the multi-modal customer care center. Our central component is a VoiceXML based voice portal, into which we integrate additional components for audio and video processing. The architecture is shown in Figure 1.

Real-time Avatar Animation System

Tracking Human faces and body extremities [8] offers a possibility to transfer motions from a Human to an artificially created character or to record movements, e.g. for different parts of the dialog, with the goal of animating the character using a combination of learned motion patterns and features generated on-line. Within our avatar system, avatars can be created by choosing from a given inventory (head, clothes, hair) or by "cloning" a real person, which means automatically choosing the best picks based on a photograph.

The most important components of our on-line avatar animation system are a hand gesture recognition system as presented in [5] and a facial expression estimation system [3]. In our basic setting, the position of the agent's hands and several feature points for head orientation tracking are estimated from the video signal, converted to standard MPEG-4 face and body animation parameters (FAP/ BAP), and transmitted to the customer over the network for reconstruction at the client side.

Figure 2. Example picture from user's own camera, avatar representation generated from own camera picture (bottom right), and avatar as generated from incoming FAPs/ BAPs. Avatars are shown on neutral background.

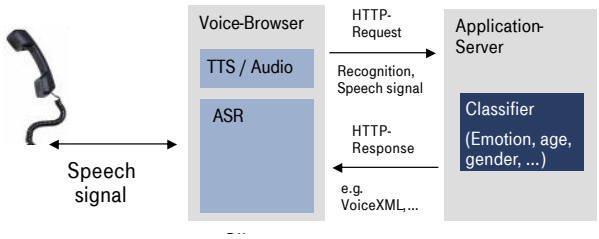

Figure 3. The technical realization of the enhanced IVR system: extra classifiers are run on the application server in order to avoid non-standard modifications to the voice browser.

More details on the algorithms used for parameter initialization, smoothing, and 2D/ 3D conversion, as well as on the fusion of audio and video information to animate lips resulting in reliable and convincing reconstruction of a moving agent at the client side can be found in [10]. Example avatars of our current system are shown in Figure 2.

Enhanced Speech Server

Virtually all automatic speech recognition systems deployed today communicate with the environment using an interface based on VoiceXML. A voice portal usually consists of a voice browser, which runs the speech recognizer, and an application server, which hosts the dialog manager and the application logic. To avoid having to perform modifications to the voice browser and to retain compatibility with VoiceXML 2.0, we integrated the modules for the detection of emotion, gender, and age as servlets on the application server, and forward the audio data to the application server as well. In the long term, integration into the voice browser would of course be preferable to reduce traffic between the servers. The system architecture is shown in Figure 3. Connections with the avatar system will be realized as close to the voice browser as possible, in order to guarantee synchronization of audio and video streams.

DETECTION OF EMOTION, GENDER, AND AGE

The second major contribution to the vision of an enhanced customer support center are the component technologies allowing the semantic interaction module (usually also run-

ning on the application server) to interpret not only the lexical transcription of what was said, but to also take into account voice-based, but non-verbal information.

Emotion Detection

The idea of automatic detection of emotions has been around for some time and has been proposed for many applications; following the terminology established in [1], our application focuses on applications in "emotional monitoring" and "believable agents".

For our experiments, we initially distinguished three degrees of emotion: (1) not angry, (2) low anger, and (3) high anger. The distinction between these two types of anger is not to be confused with the "hot" and "cold" anger labels (c.f. [11]), but is motivated by the wish to be able to detect two degrees of anger, allowing us to use different conciliation strategies for "serious" and "mild" anger [2]. The idea is that "slightly" angry callers can be calmed by more conservative conciliation strategies than those that are downright furious, a distinction we can only describe as "degree of anger". We hoped it would express itself in a difference with respect to the acoustic manifestations of the anger. However, this distinction could not be used successfully in our experiments, so that it will not be retained in future experiments.

To detect anger, we distinguish three types of features:

Prosodic features derived from the acoustic signal: our current approach is based on 31 features derived from pitch and energy, computed as one feature vector for a whole utterance. Gaussian classifiers were trained on a 3h training set taken from a real dialog system prototype [2]. Reference labels were generated by three Human labellers experienced in development and quality monitoring of Man-Machine dialog systems. Only sections which all three labellers marked as "angry" were treated as such. Inter-labeller agreement however was low: while two labellers agreed fairly well (80%), a third one disagreed quite often.

Lexical features given by spotting of words from a manually determined swear-word list. The use of lexical machine learning methods like e.g. the ones described in [6] was prevented by the restricted dialog grammar.

Dialog History features were computed by calculating prosodic deviation from non-angry speech with respect to the first turn uttered by the user and also by counting the detected anger occurrences in the dialog and lowering the threshold if anger appeared repeatedly. As shown in Table 1 and discussed in [2], this approach allowed us to distinguish low from high anger at least to some degree.

The overall decision is generated by comparing confidences generated by class-specific one-vs-all detectors. Test data consists of 26 min. of data from the same sources as the training data.

Classifier training relies on consistently labeled data. As the decision on whether callers are angry based on short, distorted utterances from a voice portal is difficult for Humans, too, fabrication of such data is a non-trivial process. We are therefore participating in the W3C's Emotion Incubator Group[1], which deals with a standardized emotion markup language. In our setup, we managed to adjust thresholds carefully and achieve reasonable anger recall without too many false alarms, which in our experience result in the comical effect of the system trying to "calm" a perfectly "happy" customer.

Truth	Classified as		
	No anger	Low anger	High anger
No anger	89%	9%	1%
Low anger	46%	49%	4%
High anger	28%	54%	16%

Table 1. Confusion matrix for automatic anger detection on a database of 26 min. of speech (6 min. containing anger).

Age and Gender Recognition

Knowledge about age and gender of a caller can be exploited in several ways, particularly if the client group expected to call a certain service is inhomogeneous and groups of different age and/ or gender generally have different expectations or different previous experience. This is the case for example for cinema booking hotlines, recommendation systems, or for some technical hotlines. A classic example is generally a tailored response to special needs of elderly people [9]; in this work, however, we plan to introduce specialized variants of our portals for other groups as well.

In this project, we compared several approaches to age and gender detection on short (<6 sec. average) segments of telephone quality speech, for which reference information was available. Results show that Human and machine performance are not too far apart [7], although hard decisions on age, for example to control access to specific services for minors, are not possible. Still, our results show that 7 age/ gender classes[2] can be identified with approximately 55% precision and recall, for example by using a system based on parallel phone recognizers for the individual age groups. As most confusions occur within neighboring age groups and the male/ female distinction generally works more robustly, these classifiers are usable in practice, as dedicated interaction strategies may not be available for all 7 age/ gender classes anyway, but only for men and women or young people and elderly people. We are currently analyzing existing data to develop guidelines for the development of class-specific interaction models.

EVALUATION AND USER STUDIES

For several prototype systems modeling sub-domains of production IVR systems, internal usability and acceptance tests were carried out, from which we draw the following main observations:

[1] http://www.w3.org/2005/Incubator/emotion/

[2] These were "Children" plus "young", "adult", and "senior" for "Males" and "Females". These classes were chosen so that different application scenarios initially envisaged could be tested by collapsing classes, i.e. without having to retrain the classifiers for every scenario.

User Experience with Avatars

When testing three scenarios loosely based on "chat"-type applications with a total of 46 users, we found the acceptance of this technology to be highly dependent on the overall functionality and usability (e.g the consistency of the GUI containing the avatar window etc.) of the service. Acceptance was higher among younger people, which did even pronounce avatars to be "fun". The current output technology eventually raises expectations, which can hardly be met with present sensor capabilities, e.g. reliable recognition of mimics. The solutions raise expectations on a huge variety of "peripheral" features: more avatars, more accessory, option to include background videos in avatar messages etc.

Emotion Detection

In an internal acceptability study, about a third of 200 users of a prototype "emotion-aware" IVR dialog system reported afterward that they had spoken "angrily" to the system while dealing with one of 5 representative tasks. As we had used conservative settings for the system reaction on assumed "angry" input, about a fifth of these users said they had noticed a change in system behavior as a reaction to their angry speech, which 70% of users appreciated. Our users generally liked the idea of an emotion-aware voice portal significantly more, once they had experienced targeted system behavior. This observation seems particularly useful for infrequent callers, who may not be familiar with the optimal strategy to deal with automatic IVR systems.

Recognition of Age and Gender

User studies using controled conditions for age and gender of callers showed that dialog strategy and familiarity with IVR systems can vary significantly between groups such as men, women, and seniors [4]. While only applicable for heterogeneous user groups, results from a study involving 25 users who called a dialog system in typical customer care scenarios [4] indicate that adaptation of ASR and dialog components could yield significant gains in dialog success rates, usability, and ultimately customer satisfaction.

SUMMARY AND CONCLUSION

With the increasing availability of interactive TV over broadband IP networks in homes, the advent of mobile TV, and the availability of multi-modal capabilities at many other locations such as shops and other public spaces, we investigate a scenario in which customers, or users in general, interact with an automatically animated call center agent, to provide for a richer interaction than currently possible over plain switched telephony networks.

This paper summarized component evaluations and our experience with user studies and acceptability tests on the use of avatars, emotion detection, and recognition of age and gender, which encourage us to also upgrade the "upstream" information flow in a customer care scenario and to "put the animated agent closer to the user". An important part of our system development efforts, which could not be discussed here, is the ability to easily and consistently derive variants of an existing dialog, which are specific to a particular user group, for example young men or elderly people. For this we have designed a "workbench" in which we are constantly integrating design patterns and transformation rules, so that in the future it should be possible to create variants of existing dialogs and processes semi-automatically.

REFERENCES

1. BATLINER, A., BURKHARDT, F., VAN BALLEGOOY, M., AND NÖTH, E. A taxonomy of applications that utilize emotional awareness. In *Proc. Fifth Slovenian and First International Language Technologies Conference* (Ljubljana; Slovenia, Oct. 2006), SLTS.
2. BURKHARDT, F., VAN BALLEGOOY, M., ENGLERT, R., AND HUBER, R. An emotion-aware voice portal. In *Proc. 16. Conference for Electronic Speech Signal Processing (ESSP) 2005* (Prague; Czech Republic, Sept. 2005).
3. EISERT, P., AND GIROD, B. Analyzing facial expressions for virtual conferencing. *IEEE Computer Graphics & Applications 18*, 5 (Sept. 1998), 70–78.
4. HEMPEL, T. Usability of a telephone-based speech dialogue system as experienced by user groups of different age and background. In *Proc. 2nd ISCA/DEGA Tutorial & Research Workshop on Perceptual Quality of Systems* (Berlin; Germany, Sept. 2006), ISCA.
5. JUST, A., BERNIER, O., AND MARCEL, S. HMM and IOHMM for the recognition of mono- and bi-manual 3d hand gestures. IDIAP-RR 39, IDIAP, 2004.
6. LEE, C. M., AND NARAYANAN, S. S. Toward detecting emotions in spoken dialogs. *IEEE Transactions on Speech and Audio Processing* (2005).
7. METZE, F., AJMERA, J., ENGLERT, R., BUB, U., BURKHARDT, F., STEGMANN, J., MÜLLER, C., HUBER, R., ANDRASSY, B., BAUER, J. G., AND LITTEL, B. Comparison of four approaches to age and gender recognition for telephone applications. In *Proc. ICASSP 2007* (Honolulu, Hawaii, Apr. 2007), IEEE.
8. MOESLUND, T. B., AND GRANUM, E. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU 81*, 3 (2001), 231–268.
9. MÜLLER, C., WITTIG, F., AND BAUS, J. Exploiting speech for recognizing elderly users to respond to their special needs. In *Proc. Interspeech 2003 (Eurospeech)* (Geneva; Switzerland, Sept. 2003), ISCA.
10. SCHREER, O., TANGER, R., EISERT, P., KAUFF, P., KASPAR, B., AND ENGLERT, R. Real-time avatar animation steered by live body motion. In *ICIAP* (2005), F. Roli and S. Vitulano, Eds., vol. 3617 of *Lecture Notes in Computer Science*, Springer, pp. 147–154.
11. YACOUB, S., SIMSKE, S., LIN, X., AND BURNS, J. Recognition of emotions in interactive voice response systems. In *Proc. Interspeech 2003 (Eurospeech)* (Geneva; Switzerland, Sept. 2003), ISCA.