S. Boll, S. Maaß & R. Malaka (Hrsg.): Workshopband Mensch & Computer 2013
München: Oldenbourg Verlag, 2013, S. 531–534

CARUSO – Singen wie ein Tenor

Jochen Feitsch, Marco Strobel, Christian Geiger

FH Düsseldorf, Fachbereich 5 - Medien, Mixed Reality & Visualisierung

Zusammenfassung

In diesem Beitrag beschreiben wir ein Projekt, das darauf abzielt dem Benutzer das Gefühl zu geben, wie ein Tenor zu singen. Wir kombinieren 3D-Ganzkörpertracking mit Gesichtstracking, Morphing, Gesangssynthese und 3D-Character Rendering in einer interaktiven Medieninstallation.

1 Einleitung und Übersicht verwandter Arbeiten

Ziel des Projektes ist eine interaktive Medieninstallation, die dem Benutzer die Möglichkeit bietet, sich wie ein Tenor aus dem 20. Jahrhundert zu fühlen und als solcher eine Arie zu singen. Bei der Entwicklung einer solchen musikalischen Benutzungsschnittstelle sind das Verfolgen von Körpergesten, die Erkennung der Mundformung sowie eine geeignete Sound- und Vibrationsausgabe wichtige Komponenten. Damit es dem Nutzer möglich ist, sich in Grenzen wie ein Tenor zu verhalten und zu fühlen, sollen unterschiedliche Interaktionstechniken eingesetzt werden, die visuelles, akustisches und haptisches Feedback für eine glaubwürdige Benutzererfahrung bereitstellen. Die Generierung der Gesangsstimme erfolgt rein synthetisch durch optisches Tracking der Mundöffnung. Dem Benutzer wird die Möglichkeit gegeben, sowohl die gesungene Tonhöhe, als auch den gesungenen Vokal jederzeit zu variieren. Dies geschieht über Heben und Senken der Arme, bzw. über das Formen des gewünschten Vokals mit dem Mund.

Andere Projekte haben bereits die Synthese von Sound mit Mundgesten untersucht. De Silva et al. [7] zeigten eine Mundsteuerung mittels Gesichtserkennung. Die bildbasierte Erkennung identifizierte die Nasenloch- und Mundform und übertrug diese Parameter auf das Gesangsmodell. Lyons et al [5] präsentierten ein visuelles Mund-Interface, das Gesichtsbewegungen nutzte um musikalische Klangereignisse zu generieren. Eine am Kopf getragene Kamera verfolgte Mundhöhe und -breite sowie das Verhältnis dieser Parameter. Diese wurden weiterverarbeitet um Gitarreneffekte bzw. ein Keyboard zu steuern. Vergleichbar zu unserem Ansatz ist das “Artificial Singing“ Projekt, bei dem ein Gesangssynthesizer mit Mundbewegung gesteuert wird, die von einer WebCam aufgenommen wurden [8]. Der Mund des Benutzers wird dabei verfolgt und Parameter wie Breite, Höhe, Öffnungsgrad, Mundform, etc. werden auf die Synthesizerparameter wie Tonhöhe, Lautstärke und Vibrato etc. übertragen. Cheng und Huang haben einen fortgeschrittenen Ansatz veröffentlicht, der Mundtracking und 3D Rekonstruktion in Echtzeit bietet [6]. Das Synthetisieren von Gesang ist ein ehrgeiziges Forschungsgebiet mit langer Tradition, da die menschliche Gesangsstimme ein sehr komplex zu synthetisierendes Instrument ist. Eine gute Übersicht bieten die Publikationen [9, 10]. Das

in unserem Fall notwendige Erstellen und Steuern eines 3D Avatars wurde ebenfalls bereits in einigen Projekten diskutiert. FaceGen (www.facegen.com) ist eine bekannte Basistechnologie, die in einigen aufwändigen 3D-Computerspielen genutzt wird. Viele Musikinterfaces, wie auch unseres, setzen RGB-D Kameras wie Microsoft Kinect, Asus XtionPro oder PrimeSense Produkte ein, um die Soundsynthese zu steuern (z.B. [11]).

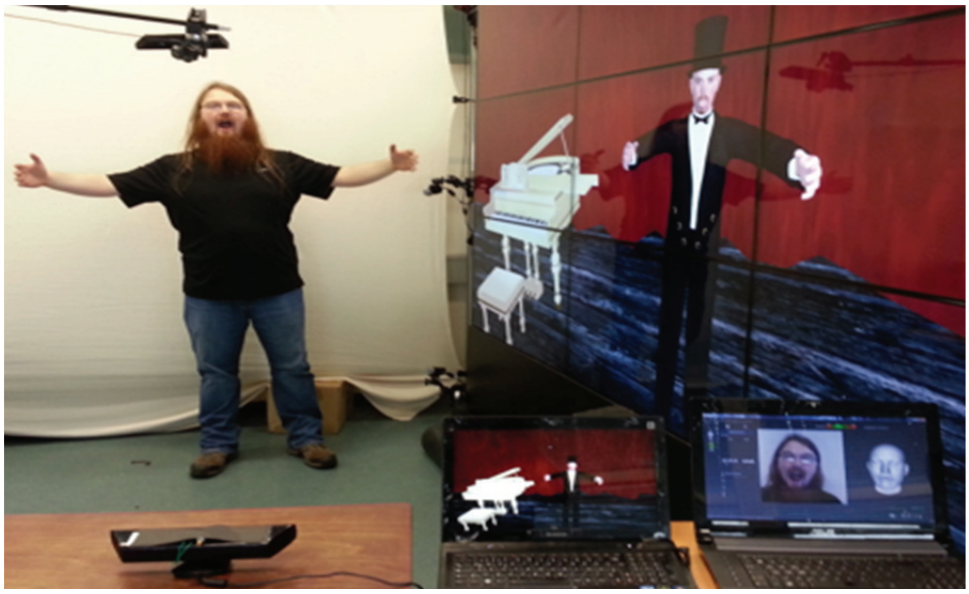

Abbildung 1: Beispielhafter Aufbau der Installation

Die Installation besteht aus einer 3x3 Videowand mit 46“ Monitoren, einer Microsoft Kinect und wahlweise einem Primesense Carmine oder einer weiteren Microsoft Kinect. Der Benutzer steuert einen virtuellen 3D-Tenor in einer großen Opernkulisse. Das Gesicht des Tenormodells kann zu Beginn vom Benutzer angepasst werden. Durch Bewegen der Arme und das Formen von Vokalen mit dem Mund kann der Benutzer die Bewegung des Tenors steuern und diesen zum Singen bringen und dadurch selbst in die Rolle des Tenors schlüpfen.

2 Benutzertracking, 3D-Rendering, Hardware

Das Tracking nutzt die Skelettdaten der Microsoft Kinect und verwendet darüber hinaus auch ein Kopf- und Gesichtserkennungssystem: faceshift (http://www.faceshift.com) berechnet die benötigten Daten aus Tiefen- und RGB-Bildern und stellt diese der Anwendung zur Verfügung. Die aktuelle Mundform des Benutzers wird durch Attribute wie Öffnungsgrad in Höhe und Breite und weiterer Daten zur Mundform analysiert. Diese Informationen werden an die Klangsynthese gesendet und zur Erzeugung des Gesangs genutzt. Handposition und Armstreckung des Benutzers dienen zur Veränderung der Tonhöhe und Lautstärke. Je weiter der Arm gestreckt ist, desto lauter wird der Ton. Die Position der Hände in Relation zu den Schultern des Benutzers dient zum Steuern der gesungenen Tonhöhe. Je tiefer die Hände desto tiefer ist der erzeugte Ton. Je höher die Hände gehalten werden, umso höher ist der

Ton. Dabei betrachtet das System jeweils beide Hände, verarbeitet jedoch nur die Werte der höher positionierten Hand. Die grafische Darstellung wird über die Unity3D Game-Engine erzeugt. Bei Start der Installation erstellt der Benutzer zunächst ein individuelles 3D-Model seines Gesichts. Dies geschieht mittels einer HD Kamera und einer Implementierung des FaceGen SDK, bei der ein 2D-Foto auf ein 3D-Polygon Mesh abgebildet und die Kopfform anpasst wird, um eine möglichst gute Rekonstruktion des Benutzers zu erzielen. Der erstellte Kopf wird dann auf den Körper eines vormodellierten 3D-Tenors gesetzt. Über den Microsoft Kinect Sensor wird die Bewegung des Nutzers auf den Körper des Avatars und die Daten der Gesichtserkennung auf dessen Mundform und andere Gesichtsmerkmale übertragen. Das System bietet eine Visualisierung, die anzeigt, wie gut der User die gewählte Arie präsentiert. Dabei verändert sich der virtuelle Kopf langsam vom Benutzerkopf zu einem 3D-Modell des Tenors Enrico Caruso.

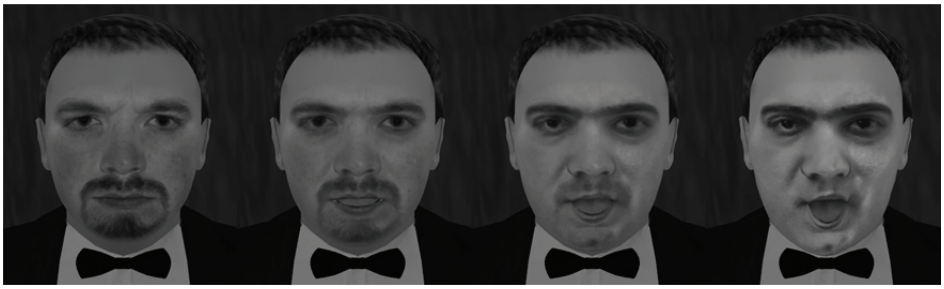

Abbildung 2: Gesicht des Benutzers morpht zu Gesicht von Enrico Caruso – aus [12]

3 Klangsynthese und Mapping

Die Klangsynthese geschieht mittels Vokalsynthese in Max/MSP (www.cycling74.com), genauer mittels Formantsynthese. Die vom Benutzer durch die Armposition bestimmte Tonhöhe dient als Grundfrequenz, zu der zunächst mehrere Sinuswellen zu einem Grundsignal addiert werden. Diese addierten Sinuswellen weisen hierbei jeweils eine Frequenz auf, die ein ganzzahliges Vielfaches der Grundfrequenz ist. Die Obergrenze für die so erzeugten „Obertöne“ des Grundsignals wurde auf 12kHz festgelegt, um das Frequenzspektrum der menschlichen Stimme zu imitieren. Das so erhaltene Signal wird anschließend mehrfach moduliert. Die verwendeten Modulationsfrequenzen basieren hierbei ebenfalls auf der Grundfrequenz, werden allerdings selbst ständig durch randomisierte Signale moduliert um ein möglichst realistisches Verhalten des Grundsignals zu erreichen. Zusätzlich wird durch die Modulation ein Vibrato-Effekt und eine leichte, ebenfalls zufällige Schwankung der Signalamplitude hinzugefügt, um das Verhalten einer menschlichen Gesangsstimme zu simulieren. Das so erhaltende Grundsignal wird daraufhin parallel durch vier Bandpass-Filter gesendet, die die charakteristischen Formanten des gewünschten Vokals erzeugen, und gewichtet wieder zusammengeführt. Mehrere zusätzliche Filter reduzieren unangenehme Frequenzanteile und sorgen für einen Klang wie aus einem Grammophon des 19. Jahrhunderts. Abschließend werden vokalabhängige Amplitudenunterschiede durch Normalisierung des Signals und einen Kompressor ausgeglichen. Optional kann noch ein Hall-Effekt hinzugefügt werden. Der Benutzer steuert die Lautstärke und die Tonhöhe des Gesangs mit den Armen. Die Lautstärke wird hierbei durch die Streckung der Arme direkt auf die Lautstärke gemappt. Die Armhöhe wiederum steuert die Tonhöhe und wird auf 25 Stufen quantisiert, von MIDI-

Pitch 41 (entspricht circa 87 Hz) bis MIDI-Pitch 65 (entspricht circa 349 Hz). Diese 25 Stufen werden wiederum auf die aktuelle Tonleiter gemappt. Hierzu wird über eine Tabelle der entsprechend nächste zur Tonleiter passende MIDI-Pitch zugeordnet. Zugrunde liegen hierbei verschiedene Tonleitern basierend auf dem Grundton C, jeweils in mehreren Variationen (zum Beispiel mit hinzugefügter Septime). Das Mapping unterscheidet, ob die aktuelle Zählzeit auf 1 bzw. 4 (ausgehend von einem 4/4 Takt), oder 2 bzw. 3 ist. Im ersten Fall wird nur auf Töne des entsprechenden Drei- bzw. Vierklangs gemappt, im zweiten Fall ist ein Ansteuern aller Töne der Tonleiter möglich. Hierdurch soll eine möglichst wohlklingende Melodieführung erreicht werden. Um andere Grundtöne als C zu realisieren, wird die entsprechende Tabelle transponiert. Übersteigt hierbei einer der MIDI-Pitches den Wert 65, wird dieser unten (beginnend bei MIDI-Pitch 41) in der Tabelle ergänzt und die Tabelle neu sortiert. Der gewünschte Vokal wird über die Mundöffnung des Benutzers bestimmt. Hierzu werden drei Parameter betrachtet, die als aussagekräftig für die Vokale „A“, „E“ und „O“ identifiziert wurden. Der jeweils stärkste ausgeprägte Wert bestimmt dabei den gesungenen Vokal. Zusätzlich kann der Mund geschlossen werden, um nicht zu singen. Konkret werden durch die verschiedenen Mundstellungen die Frequenzen der bei der Klangsynthese verwendeten Formanten und deren Gewichtungen bestimmt.

4 Fazit

Der aktuelle Projektstand ist ein funktionsfähiger Prototypen, der allerdings noch eingeschränkt ist durch die Tracking-Qualität und das Spektrum der synthetisierten Vokale. Für eine glaubwürdige Simulation der Gesangsstimme arbeiten wir an zusätzlichem Feedback für den Benutzer durch Vibration des Torsos. Des Weiteren testen wir Knochenschall-Kopfhörer, um den Gesang „im Kopf“ des Benutzer zu platzieren.

REFERENCES

5 Lyons, M., Haehnel, M. and Tetsutani, N. 2003 Designing, Playing, and Performing with a Vision-based Mouth Interface, In *Proc. of NIME 2003, Montreal, Canada.*

6 Cheng, J. and Huang, P. 2010. Real-Time Mouth Tracking and 3D reconstruction. *3rd Int. Conf on Image and Signal Processing*.

7 de Silva, C., Smyth, T., and Lyons, M. J. 2004. A novel face-tracking mouth controller and its application to interacting with bioacoustic models. In *Proceedings of NIME 2004,* Hamamatsu, Japan.

8 Hapipis, A. and Miranda, E. R. 2005. Artificial Singing with a webcam mouth-controller. *Int Conf. on Sound and Music Computing*, Salerno, Italy.

9 Rodet, X. 2002. Synthesis and Processing of the Singing Voice. *1st IEEE Workshop on Model Based Processing and Coding of Audio*, Leuven, Belgium.

10 Sundberg, J. 2006. The KTH Synthesis of Singing. *J. on Advances in Cognitive Psychology.* Vol 2, No 2-3.

11 Odowichuk, G., Trail, S., Driessen, P., Nie, W., and Page, W. 2011. Sensor fusion: Towards a fully expressive 3D music control interface. *IEEE Pacific Rim Conf. on Communications, Computers and Signal Processing*

12 Feitsch, J., Strobel, M., Geiger, C., CARUSO – Augmenting Users with a Tenor's Voice. Augmented Humans, AH’13, Stuttgart, 201