Proseminar Gesprochene Sprache

Emotionen in Sprachsynthesesystemen (draft)

B.sc. Informatik Sommersemester 2020

Jan Goldmann

Abstract

Als Emotionen versteht man einen inneren Zustand der durch Ereignisse oder Situationen ausgelöst wird. Die Vermittlung dieser inneren Zustände über die Sprache ist ein wesentliches Merkmal der menschlichen Sprache. Seit Jahren wird untersucht wie Emotionen in Sprache transportiert werden kann, und welchen Einfluss sie auf die Zuhörer*innen hat. Für Sprachsynthese Systeme, die einen menschlichen Klang anstreben scheint langfristig eine glaubhafte Darstellung von Emotionen unerlässlich zu sein. Diese Arbeit soll zeigen auf welchen Weisen dieses Problem angegangen wurde und welche Schwierigkeiten immer noch bestehen.

Inhaltsverzeichnis

1 Einleitung

1.1 Anwendung - Die Stimme der Maschinen.

Elektronische Sprachsysteme, meistens als Text-to-speech (im folgenden: tts) Programme stellen eine immer wichtigere Schnittstelle zwischen Maschinen und Menschen dar. Der Umgang mit elektronischen Hilfsgeräten soll immer einfacher und natürlicher ablaufen, dabei ist wird eine menschlichere Kommunkationsform wie Sprechen, die visuelle Kommunikation in vielen Bereichen ersetzen. Aber auch für die Kommunikation zwischen Menschen sind tts-Systeme in Form von Sprachassistenten sowohl für Menschen mit körperlichen/geistigen Einschränkungen als auch in Übersetzungsprogrammen relevant. Seit Jahren wird versucht diese Systeme möglichst menschlich klingen zu lassen [8]. Es hat sich gezeigt, dass durch eine emotionale Einfärbung der synthetisierten Stimme die Verständlichkeit des Inhaltes stark steigt. Für die Anwendung im prosthetischen Bereich ist die Möglichkeit Emotionen über Klang zu vermitteln sehr wichtig. Es gibt also vielseitige Gründe tts-Systeme Emotionen transportieren zu lassen.

1.2 Sprachsynthese - kurze Einführung

Bei tts-Systemen wird aus dem eingegebenen Text über viele Schritte eine Klangfolge generiert die bestenfalls als menschlich wahrgenommen wird, jedoch mindestens den Anspruch hat den eingegebenen Inhalt verständlich wiederzugeben. Um das zu erreichen wird der eingebene Text in Satzbausteine zerlegt, dabei werden regelbasiert Abkkürzungen Zahlen und Satzzeichen gelabelt. Nebensätze und Hauptsätze werden erkannt und markiert. Anschliessend die Endung angepasst und die einzelnen Wörter und Phoneme mit markern versehen die die Betonung angeben. An dieser Stelle können Prosodieregeln angewandt werden, die mit Emotionen zusammenhängen. Dabei kann auf eine Datenbank zugegriffen werden oder auf ein Set von Regeln, die die Ausprache beeinflussen. Diese markierung werden akustische Paramter übersetzt die letzendlich die tatsächliche Lauterzeugung beeinflussen.

2 Emotionen in Sprache

2.1 Emotionen erzeugen - unterschiedliche Parameter und Methoden

Gesprochene Sprache weist viele Eigenschaften auf, welche man getrennt betrachten kann. Das Klangspektrum, Prosodie und Phonetik sind dabei die Eigenschaften die den Klang betreffen, während Wortwahl und Semantik den Inhalt betreffen. Da in TTS-Systemen der Text, also der Inhalt, übergeben wird, bleiben nur die klanglichen Merkmale um Emotionen darzustellen. Je nach Angewandtem Sprachsynthese Modul lassen sich diese Parameter unterschiedlich stark beeinflussen. Die Formant synthese ist dabei die Methode bei der die größte Kontrolle über die Sprachsynthese besteht, da dabei keine Aufnahmene von Menschen verwendet werden, sondern die Sprache direkt von der Maschine erstellt wird. Entsprechend klingt Sprache die mit der Formant synthese geschaffen wurde nicht sehr menschlich. Eine Technik die schon natürlicher klingt ist die Diphone Konkatenation. Die Veränderung der Diphone beschränkt sich dabei aber nur auf die Grundfrequenz,

die Dauer und Intensität. Durch die Verwendung von Unit-Selection kann das als natürlichstes wahrgenommene Ergebnis erzielt werden. Dabei ist, besteht aber eine große Abhängigkeit der verwendeten Sprachbausteine. Ein Effekt, der bei der Auswahl des Synthesesystems berücksichtigt werden sollte ist die sogenannte Akzeptanzlücke (uncanny valley), dabei steigt die Akzeptanz eines Systems nicht notwendigerweise mit dessen Natürlichkeit. Ein eher künstlich klingendes System ist daher einem natürlicherem eventuell vorzuziehen.

2.2 Emotionen erkennen - Die Maschine in der Menschenwelt

Bevor Emotionen von TTS-Systemen dargestellt werden können, müssen die unterschiedlichen Emotionen von einander abgegrenzt werden. in früheren Arbeiten wurde versucht, wenige aber dafür sehr starke, Grundemotionen darzustellen [1,2]. In vielen Arbeiten wurde Freude, Ärger, Trauer, Wut, Angst, Langeweile und Überraschung als Basis Emotionen verstanden aus denen sich alle weiteren Emotionen ableiten lassen. Diese Kategorien lassen sich aber auch in Subkategorien einteilen. Die Emotion Freude hätte dann Beispielsweise die Untertypen Stolz, Zufriedenheit und Vergnügen. Der Vorteil dieser Einteilung ist, dass sie leicht verständlich sind und in der menschlichen Kommunikation etabliert.
In der Arbeit von Cowie wird ein alternativer Ansatz verfolgt [3, 4]. Dabei werde „schwächere" Emotionen in den Vordergrund gerückt. Emotionen werden hier durch Verortung in einem mehrdimensionalen Raum dargestellt. Durch die Verknüpfung der einzelnen Dimensionen mit skalierbaren Parametern der Klangeigenschaften lässt sich so nicht nur ein genaueres Bild der einzelnen Emotionen zeichnen, sondern auch die Übergänge zwischen Ihnen lassen sich erfassen. Es wurden drei Dimensionen herausgearbeitet, welche die Emotionalität der Sprache erfassen sollen. Activation könnte man als Belebtheit oder Erregtheit der Sprache bezeichnen, während Evaluation die Zufriedenheit oder Vergnügen ausdrückt. Power als Kraft oder Dominanz ist die dritte Dimension. Activation ist die am stärksten beeinflussbare Dimension, sie lässt sich stark über die Grundfrequenz, die durchschnittliche Intensität und das Sprechtempo steuern.
Power wird über ähnliche Parameter bestimmt, dabei sind jedoch teilweise die Korrelationen umgekehrt und die Vokale Länge scheint eine größere Rolle zu spielen [1].
Evaluation wird durch längere Vokale und eine geringere Intensitätssteigerung innerhalb des Sprachsamples erreicht[1].
Durch die Einordnung unterschiedlicher Sprachsamples in diesem drei dimensionalen Raums lassen sich Prosodieregeln erstellen die für TTS benutzt werden können (Abb.1).

	Prosodic parameter	Coefficients		
		Activation	Evaluation	Power
fundamental frequency	pitch	0.3	0.1	-0.1
	pitch-dynamics	0.3%		-0.3%
	range	0.4		
	range-dynamics	1.2%		0.4%
	accent-prominence	0.5%	-0.5%	
	preferred-accent-shape		E≤-20: falling -20<E≤40: rising E>40: alternating	
	accent-slope	1%	-0.5%	
	preferred-boundary-type			P≤0: high P>0: low
tempo	rate	0.5%	0.2%	
	number-of-pauses	0.7%		
	pause-duration	-0.2%		
	vowel-duration		0.3%	0.3%
	nasal-duration		0.3%	0.3%
	liquid-duration		0.3%	0.3%
	plosive-duration	0.5%	-0.3%	
	fricative-duration	0.5%	-0.3%	
	volume	0.33%		

Abbildung 1: Prosodieregelen die Faktoren und Werte für die einzelnen Dimensionen darstellen [6].

2.3 MaryTTS und EmoSpeak - Module

in sich immer weiter entwickelndes Projekt, das mit Prosodieregeln arbeitet, ist MARY (Modular Architecture for Research on speech sYnthesis) [7] und das damit erstellte EMOSpeak Interface (Abb. 2). Dabei wurde eine Annotation (MARYXML) geschaffen die den Eingabetext mit zusätzlicher Information versieht die Einstellungen an das TTS-System übergibt. MARY funktioniert dabei modular und kann an unterschiedliche Synthese Systeme angeschlossen werden. Dabei werden dann je nach angestrebter Emotionsdarstellung die Parameter in der Sprachgenerierung, den Möglichkeiten des jeweiligen Systems, angepasst. Durch das einfache Interface ist eine intuitives Verständnis möglich. Man kann aber auch die einzelnen Schritte der Sprachsynthese einsehen und bearbeiten und kann so ein besseres Verständniss von emotion in Sprachsynthese als auch über emotion in Sprache generell erlangen.

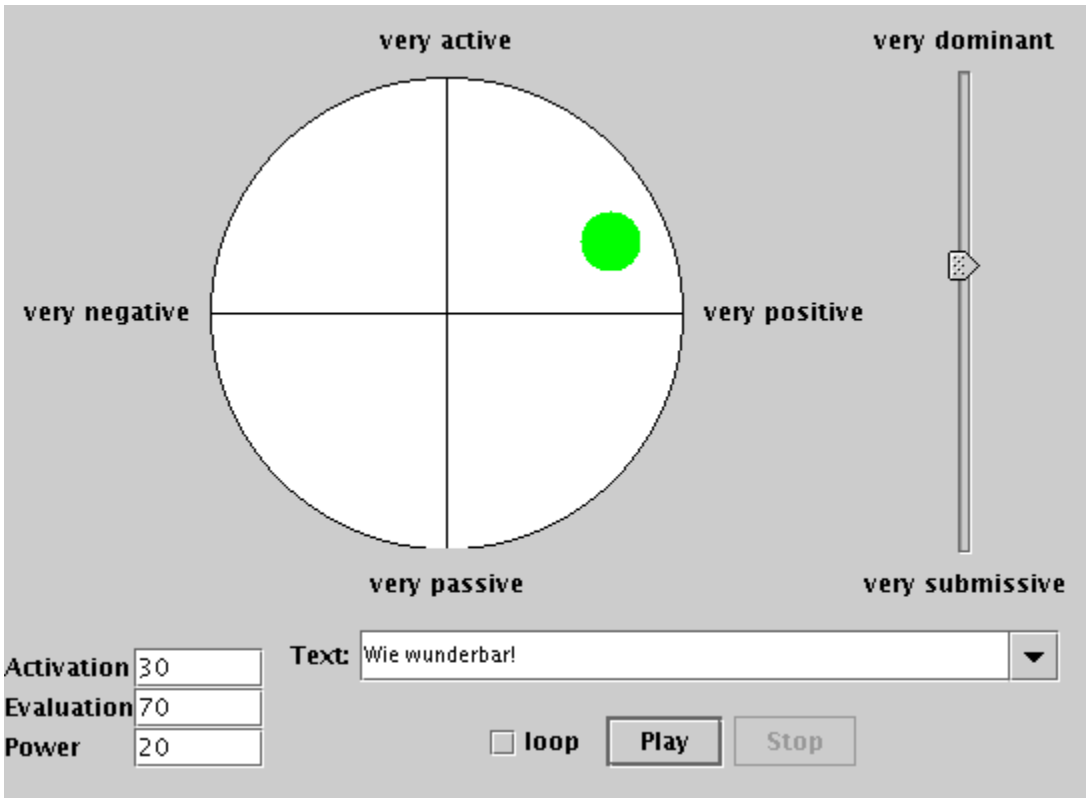

Abbildung 2: Emospeakinterface, mit Texteingabe und Reglern über die man die Dimensionsgrößen einstellen kann [6].

2.4 Evaluation - Anspruch und Wirklichkeit

Ob ein TTS-System in der Lage ist eine gegebene Emotion zu vermitteln gibt es unterschiedliche Methoden der Auswertung. Einerseits kann man das FEELTRACE-System nutzen [5]. Wird der wahrgenommene emotionale Zustand auf in den drei oben erwähnten Dimensionen eingeordnet. Debei werden jedoch weitere Träger für Emotionen vernachlässigt, z.b.: Wortwahl und Gesichtsausdruck. Bei forced choice test wird ein Text neutralen Inhalts mit unterschiedlichen Emotionen synthetisiert und muss dann nach dem entweder oder Prinzip zugeordnet werden. Ein

anderer Weg ist es bekannte Textbeispiele, bei denen die Emotionalität bekannt ist über das TTS-System sprechen zu lassen und zu überprüfen ob der Inhalt des Textes zum Klang passt. Durch eine detaillierte Auswertung lassen sich nicht nur die Parameter der einzelnen emotionalen Zustände besser definieren auch die Unterschiede oder Ähnlichkeiten lassen sich erfassen und schaffen so ein generelles Verständnis von Emotionen in Sprache.

3 Fazit

Bei der Darstellung von Emotionen in gesprochener Sprache über TTS gibt es noch vielfältige Probleme. Es werden nicht alle Emotionen gleichermaßen gut transportiert. Es ist möglich, viele Emotionen auf unterschiedliche weisen vermittelt werden können. Unterschiedliche Sprecher*innen haben unterschiedliche Methoden um gleiche Emotionen auszudrücken [9]. Andere Aspekte der menschlichen Sprache sind bei TTS auch schwer zu berücksichtigen. Menschliche Sprache besteht auch zu großen Teilen aus nicht verbalen Vokalisation. Dabei werden beim Sprechen unterschiedliche Geräusche gemacht, die keinen direkten Inhalt transportieren aber der Sprache eine Natürlichkeit verleihen, die nur vorgelesenem Inhalt fehlt [10]. Der Gesichtsausdruck während des Redens spiegelt nicht nur Emotionen wider, sondern hat auch Einfluss auf den Klang der Sprache. Durch Übertreibung der Darstellung der Emotionen kann man Versuchen das Fehlen dieser weiteren Informationen auszugleichen.
Die Anpassung des Inhalts an die zu transportierende Emotion ist eine weitere Schwierigkeit. Da für den Computer der Inhalt und die Form in keinem direkten Zusammenhang stehen, muss der Text oder der Klang immer explizit aneinander angepasst werden. Eine Weiterentwicklung von Text-to-Speech hin zu concept-to-speech könnte sich in der Zukunft abzeichnen. Dabei ist der Computer selber in der Lage, zu einer abstrakter gehaltenen Information, sowohl den passenden Inhalt als auch die entsprechende Emotionalität zu finden und auszudrücken.

4 Quellenverzeichnis.

[1] Schröder, M.: Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis. PhD thesis, Institute of Phonetics, Saarland University (to appear)

[2]. Schröder, M.: Emotional speech synthesis: A review. In: Proceedings of Eurospeech 2001. Volume 1., Aalborg, Denmark (2001) 561–564

[3] Cowie, R.: Describing the emotional states expressed in speech. In: Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland (2000) 11–18

[4] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.: Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18 (2001) 32–80

[5]. Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schr¨oder, M.: 'FEELTRACE': An instrument for recording perceived emotion in real time. In: Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland (2000) 19–24

[6] Schröder, M.: Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In: Workshop on Affective Dialogue Systems (2004)

[7] Schr¨oder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology 6 (2003) 365–377

[8] Schröder, Marc. (2001). Emotional speech synthesis: a review.. 561-564.

[9] Schröder, M., Can emotions be synthesized without controlling voice quality?, *Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland*, p. 37-55. http://www.dfki.de/~schroed

[10]Trouvain, J. & Truong, K. 2012. Comparing non-verbal vocalisations in conversational speech corpora. Proc. 4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals, Istanbul, pp. 36-39.