



What do people hear? Listener's Perception of Conversational Speech

Präsentation von:

Emily Baser, Laura Jaudt, Sophie Vasen



Untersuchungsansatz:

- ◇ Pilotstudie: wie nehmen Menschen gesprochene Sprache wahr?
- ◇ Text-to-speech vs. natürliche Sprachbeispiele
- ◇ Bewertung anhand von Kriterien wie Verständlichkeit und Natürlichkeit
- ◇ Fokus auf Prosodie
 - Unterscheidung zwischen gesprochener und vorgelesener Sprache



Zielsetzung:

- ❖ Bedarf an besseren TTS Systemen vergrößern, weil Konversationsagenten immer beliebter werden
- ❖ Besser verstehen, welche Merkmale für Zuhören entscheidend sind



Forschungslücke:

- ◇ Bisherigen TTS-Modelle scheitern oft → Gesprächssprache klingt nicht natürlich
- ◇ Unklar, warum bestimmte TTS-Ansätze besser funktionieren als andere → wenig Wissen über Zuhörer



Ansätze beim Aufbau von Konversationstexten für TTS:

- ◇ 1. öffentliche Datensammlungen verwenden
(z.B. sponatne Sprache, Gesprächskorpora)
- ◇ 2. auf gezielt ausgewählte Stilmuster trainieren
(spontane, informelle Sprache)
- ◇ 3. direkte Zielvorgaben für Sprachsynthese



Prosodie in spontaner Sprache:

◈ Spontane Sprache unterscheidet sich durch:

- Unvollständige Sätze
- Häufige Pausen
- Schwankende Tonverläufe

→ Prosodie spielt dabei zentrale Rolle beim Signalisieren verschiedener Sachen

◆ Subjektive Bewertung von konversationellem TTS

- ◆ Angemessenheit im gegebenen Kontext
- ◆ Bewertungen mittels linearer Skalen und Vergleichen
- ◆ Rasante Fortschritte im TTS-Bereich erhöhen Dringlichkeit für neue Bewertungsparadigmen
- ◆ Formulierung der Anweisungen und Bewertung der Eignung in einer kommunikativen Situation

◊ Vorstudie:

- ◊ Verständnis der akustischen Hinweise, die Zuhörer verwenden, um Sprache in Konversationen zu identifizieren
- ◊ Interviewansätze
- ◊ Einzelgespräche mit jedem Teilnehmer
- ◊ 5 Teilnehmer hören gesprochene Stimuli, die laut abgespielt wurden
- ◊ Isolierte Äußerungen oder gruppiert als mehrere Versionen desselben Textes
- ◊ Teilnehmer konnten sich oft beide Stimuli als Konversationspartien vorstellen

- 
- ❖ Experimente zur Untersuchung, wie akustische Hinweise unabhängig vom Kontext die Beurteilung beeinflussen
 - ❖ Konversationeller Kontext:
Gesprächsverlauf in Text/Audioformat
 - ❖ Audiokontext schränkt Interpretationsspielraum ein, Antworten wurden dennoch durch eigene Vorstellung beeinflusst
 - ❖ -> Kontext hat einen signifikanten Einfluss auf die Wahrnehmung von konversationellem Stil

◆ Versuchsaufbau:

- ◆ Ziel: Verstehen, ob der konversationelle Stil allein durch die akustische Realisierung einer Äußerung vermittelt wird
- ◆ Rekrutierung von 16 englischen Muttersprachlern von der University of Edinburgh > 25-30- minütige Interviews
- ◆ Aufzeichnung der Interviews
- ◆ Einzelinterviews an offenem Ort

- 
- ❖ Aufgabe: Unterscheidung natürlicher Sprache
 - ❖ Versuchspersonen mussten unterscheiden zwischen konversationellen und nicht-konversationellen Sprachproben
 - ❖ Äußerungen entweder aus spontanen Gesprächen oder vorgelesen von einer weiblichen britischen Sprecherin (jeweils 1-4 Sekunden)

◇ Ergebnisse:

- ◇ Versuchspersonen konnten Sprachproben unterscheiden (87,3 % korrekt erkannt)
- ◇ Laut den Versuchspersonen konnte konversationelles Sprechen durch seine Flüssigkeit, Variabilität und den natürlichen Sprachfluss erkannt werden
- ◇ Rhythmus und Sprechgeschwindigkeit von großer Importanz

◇ Ergebnisse:

- ◇ Hauptaspekte gesprochener Sprache:
- ◇ Aussprache: natürliche Variation, Reduktionen und emotionale Tiefe gelten als konversationstypisch, genauso wie zu erkennbare Motivationen einer Äußerung
- ◇ Spontanität: natürlicher, ungleichmäßigerer Rhythmus; gelesener Text klingt formal und Pausen wurden mit Satzzeichen gleichgesetzt
- ◇ Prosodie: natürliche Intonationsmuster, gelesener Text klingt oft flach oder monoton

◊ Experiment 2:

- ◊ Hörbeispiele wurden von einer synthetischen Stimme vorgetragen, worüber die Versuchspersonen nicht informiert wurden
- ◊ Synthetische Proben wurden gezielt prosodisch modifiziert, was Perzeption der Hörer beeinflusst
- ◊ Variierende Intonation und wechselnder Rhythmus erhöhten die Wahrscheinlichkeit, dass Probe als Konversation wahrgenommen wurde
- ◊ Stimmklang wurde allerdings kommentiert

◆ Zusammenfassung:

- ◆ Hörer sind in der Lage gesprochene Sprache anhand akustischer Merkmale zu identifizieren
- ◆ Dabei stehen vor allem Aussprache und Rhythmus im Vordergrund, gefolgt von Prosodie
- ◆ Bei prosodischen Merkmalen hatten Versuchspersonen Probleme diese konkret zu benennen
- ◆ Diese Erkenntnisse sind wichtig, um in TTS-Systemen eine natürlichere Sprachsynthese zu ermöglichen

Vielen Dank für eure Aufmerksamkeit !