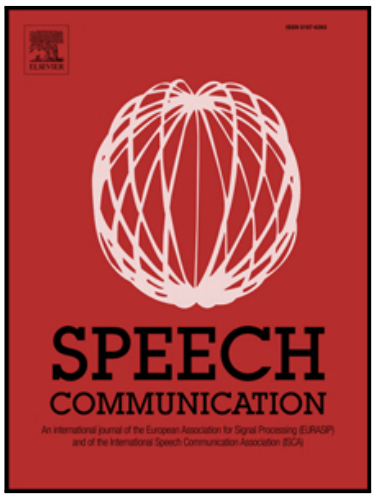


# Meaningful Head Movements Driven by Emotional Synthetic Speech

Najmeh Sadoughi, Yang Liu, Carlos Busso

PII: S0167-6393(17)30005-5
DOI: 10.1016/j.specom.2017.07.004
Reference: SPECOM 2475

To appear in: *Speech Communication*

Received date: 13 January 2017
Revised date: 1 June 2017
Accepted date: 28 July 2017

Please cite this article as: Najmeh Sadoughi, Yang Liu, Carlos Busso, Meaningful Head Movements Driven by Emotional Synthetic Speech, *Speech Communication* (2017), doi: 10.1016/j.specom.2017.07.004

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 00 (2017) 1–20

Speech Communi-cation

# Meaningful Head Movements Driven by Emotional Synthetic Speech

Najmeh Sadoughi, Yang Liu, Carlos Busso

*The University of Texas at Dallas*

**Abstract**

Speech-driven head movement methods are motivated by the strong coupling that exists between head movements and speech, providing an appealing solution to create behaviors that are timely synchronized with speech. This paper offers solutions for two of the problems associated with these methods. First, speech-driven methods require all the potential utterances of the *conversational agent* (CA) to be recorded, which limits their applications. Using existing *text to speech* (TTS) systems scales the applications of these methods by providing the flexibility of using text instead of pre-recorded speech. However, simply training speech-driven models with natural speech, and testing them with synthetic speech creates a mismatch affecting the performance of the system. This paper proposes a novel strategy to solve this mismatch. The proposed approach starts by creating a parallel corpus either with neutral or emotional synthetic speech timely aligned with the original speech for which we have the motion capture recordings. This parallel corpus is used to retrain the models from scratch, or adapt the models originally built with natural speech. Both subjective and objective evaluations show the effectiveness of this solution in reducing the mismatch. Second, creating head movement with speech-driven methods can disregard the meaning of the message, even when the movements are perfectly synchronized with speech. The trajectory of head movements in conversations also has a role in conveying meaning (e.g. head nods for acknowledgment). In fact, our analysis reveals that head movements under different discourse functions have distinguishable patterns. Building on the best models driven by synthetic speech, we propose to extract dialog acts directly from the text and use this information to directly constrain our models. Compared to the unconstrained model, the model generates head motion sequences that not only are closer to the statistical patterns of the original head movements, but also are perceived as more natural and appropriate.



*Keywords:* Conversational agents, head motion with meaning, speech-driven animation

## 1. Introduction

Head movements during conversation play an important role to convey verbal and non-verbal information. For instance, people use prototypical head movements, such as nods for backchannel [9], affirmation, and emphasis [36]. The rhythmic beat associated with head movements increases speech intelligibility, as head movements help to parse sentences [38]. But the role of head movements is not limited to conveying the lexical message. Head movements also convey the emotional state of the speaker [41, 7, 6]. Due to the multifaceted role of head movements in conversations, it is important to synthesize head movements for *conversational agents* (CAs) capturing the relation between head motion and verbal and non-verbal information.

Studies have shown that including head movements for CAs increases the level of perceived naturalness [4, 35], and the level of warmth and competence [52]. To generate head movements for CAs, studies usually rely on specifying rules based on the content of the message [8, 13, 41]. The key limitation of rule-based systems is the repetitiveness