Chaos, Solitons and Fractals 111 (2018) 175–179

Contents lists available at ScienceDirect

Chaos, Solitons and Fractals

Nonlinear Science, and Nonequilibrium and Complex Phenomena

journal homepage: www.elsevier.com/locate/chaos

Frontiers

Complexity matching in speech: Effects of speaking rate and naturalness

Adolfo G. Ramirez-Aristizabal[1,*], Butovens Médé[1], Christopher T. Kello

Cognitive and Information Sciences, University of California, Merced, 5200 North Lake Rd. Merced, CA 95343, United States

ARTICLE INFO

Article history:
Received 12 November 2017
Revised 4 April 2018
Accepted 12 April 2018

Keywords:
Complexity matching
Allan Factor
Speaking rate
Synthesized speech
Prosody

ABSTRACT

Recordings of speech exhibit nested clustering of peak amplitude events that reflects the hierarchical temporal structure of language. Previous studies have found variations in nested clustering to correspond with variations in prosody and social interaction. In the present study, we tested two specific dimensions of variation in speech hypothesized to have differing effects on hierarchical temporal structure: Speaking rate and naturalness. Rate was manipulated both algorithmically and experimentally, and naturalness was manipulated using synthesized speech, with sine wave speech as a comparison. Allan Factor analysis was used to quantify nested clustering of peak amplitude events in speech recordings as a function of timescale. For fast speech, nested clustering was found to shift into shorter timescales, whereas for synthesized speech, nested clustering was found to decrease in the longer timescales. Results are discussed in terms of complexity matching and its implications for how neural and perceptual processes might respond to changes in the hierarchical temporal structure of speech signals.

© 2018 The Authors. Published by Elsevier Ltd.
This is an open access article under the CC BY-NC-ND license.
(http://creativecommons.org/licenses/by-nc-nd/4.0/)

Measurements of speech and language commonly follow power laws [13]. These power laws suggest that underlying neural, behavioral, and social processes may be usefully theorized in terms of complex networks [2], because power laws are a natural consequence of their non-stationary, non-ergodic statistics [22]. A fundamental question about complex networks, as well as cognitive and social systems, is how they respond to inputs from their environments. For example, the dynamics of complex perceptual networks are responsive to their sensory inputs, and language networks are responsive to inputs from verbal interactions. The former is an example of unidirectional influence, because sensory systems do not directly affect the sensory world, only indirectly via the perception-action loop [9]. The latter is an example of bidirectional influence because participants in language interactions directly affect each other.

This view of cognitive and social systems as complex networks leads to predictions based on theories of how complex networks respond to external inputs. Specifically, West et al. [21] formulated the principle of *complexity matching*, which generally states that complex networks are most responsive to perturbations that match their own temporal complexity. Complexity is measured in terms of exponents that define power laws in network activity and input activity, and matching corresponds to similarity in the exponents characterizing the networks in question, and their environmental inputs. The original work defined network activity in terms of 1/f noise and fractal time series of events, the latter being analyzed in terms of waiting times (inter-event-intervals) τ, where $P(\tau) \sim 1/\tau^{\mu}$, and $1 < \mu < 2$ [21].

Recently, behavioral scientists have tested for complexity matching in human coordination and speech, based on the premise that human complex networks are highly adaptive [2]. Human complex networks may adapt by “bending” the statistics of their dynamics towards those of their inputs, to better match the environment and other complex networks. Matching is hypothesized to increase the response sensitivity of complex brain and behavioral networks. When inputs are power law distributed, matching manifests as a convergence in power law exponents of brain and behavioral networks towards the exponents of their inputs. Such flexibility in power law exponents would not be expected for less adaptive complex systems.

The first experiments to test for complexity matching in human behavior examined the dynamics of finger tapping [19], and pendula being swung together [15]. The tapping experiment used a fractal metronome that participants tried to follow as closely as possible. Fluctuations in inter-tap intervals exhibited 1/f noise, and

* Corresponding author.
E-mail addresses: aramirez64@ucmerced.edu (A.G. Ramirez-Aristizabal), ckello@ucmerced.edu (C.T. Kello).

[1] These two authors contributed equally to the study and manuscript.

https://doi.org/10.1016/j.chaos.2018.04.021
0960-0779/© 2018 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license. (http://creativecommons.org/licenses/by-nc-nd/4.0/)

power law exponents matched those of their fractal metronomes, i.e. unidirectional influence of the metronome on tapping. By contrast, the pendula experiment showed that power law 1/f exponents of angular fluctuations converged with each other, instead of a fixed stimulus like a metronome. The swinging of one pendulum by one participant was affected by the swinging of the other pendulum by the other participant, and vice versa, via perceptual and physical coupling, i.e. bidirectional influence. Together, these two studies provide evidence that human complexity matching can occur in response to stimuli in the environment, and also in response to human interactions.

One of the most natural kinds of human interaction is speech, which has also been found to exhibit complexity matching [1]. The authors recorded pairs of individuals having conversations about friendly topics with common ground, versus polarizing topics with conversational partners on opposite ends. They converted the speech waveform for each speaker into a series of acoustic onset events, and found inter-onset-intervals (IOIs) to be power law distributed like critical events of complex networks. Complexity matching was found not in IOI exponents, but in the power law clustering of events that reflects the hierarchical temporal structure of language. Specifically, Allan Factor (AF) functions for event series were closer together for conversational partners compared with baseline, but only for friendly topics for which speakers shared common ground. Polarizing conversations showed no detectable complexity matching, suggesting that the coupling of human complex networks depends on psychological and social factors, and possibly other factors as well.

Abney et al. [1] used the AF function to measure hierarchical temporal structure in speech waveforms recorded from conversations, over timescales of 30 ms–30 s. Variations in this range of timescales are perceptible to the human auditory system, and complexity matching suggests that auditory brain networks adapt the statistics of their dynamics to those of their acoustic inputs [5]. Given the relationship between complexity matching and psychological processes reported by Abney and colleagues, we hypothesize that hierarchical temporal structure in speech, as measured by AF functions, should be reflected in auditory experience by way of complexity matching in auditory networks. In support of this hypothesis, Kello and colleagues [14] found that the shapes of AF functions reflect at least three perceivable variations in complex acoustic signals: social interaction, prosodic variation, and musical composition. Greater nested clustering in peak amplitude events (as opposed to acoustic onset events) can be perceived as acoustic interactions among people, prosodic emphasis in speech, or metrical structure in music. These results are consistent with our working hypothesis, but they are quite general and do not inform how specific variations in AF functions relate to specific variations in perceivable features of speech, music, and other complex acoustic signals.

In the present study, we test two types of perceptual variations in speech that we predict to have differing effects on hierarchical temporal structure: Speech rate and naturalness. Previous studies have demonstrated consistent effects of speech rate on prosodic variation, the latter being shown to affect hierarchical temporal structure. For instance, Jun [12] found that more syllables are packed into fewer accentual phrases at faster versus slower speaking rates, thereby reducing variability by reducing the number of accentual phrases. Dellwo and Wagner [4] varied speech rates in English, French, and German, and found reduced variability in consonant durations for faster versus slower speaking rates. A modeling study in Mandarin indicated that the effect of speaking rate affects variability across several hierarchical levels of prosodic organization [3], consistent with a study of speaking rate in Mandarin [20]. In summary, previous studies indicate that faster speech should reduce prosodic variability across hierarchical levels, and thereby reduce hierarchical temporal structure across a wide range of timescales.

Speech naturalness is also predicted to affect hierarchical temporal structure, but in a different way compared with speaking rate. In particular, human-generated speech is predicted to have more hierarchical temporal structure compared with text-to-speech synthesis, particularly in the longer timescales. Variability in prosodic intonation and timing is difficult for text-to-speech synthesizers because they do not model the meanings of sentences or discourse contexts [23]. As a result, synthesized speech is often perceived as having flat affect compared with human-generated speech. Relatively flat affect should correspond with reduced hierarchical temporal structure in timescales on the order of a second and longer, as previously shown by Falk and Kello [8]. They measured AF functions in recordings of German-speaking mothers reading a story or singing a song, either to their infants or to other adults. The exaggerated prosody of infant-directed speech resulted in generally steeper AF functions, but the authors did not report a more fine-grained analysis. With respect to naturalness, Kello et al. [14] showed that AF functions for synthesized speech were flatter than those for natural speech, but again, the authors did not quantify the effect, nor did they compare it with speaking rate.

1. Allan factor analyses of speaking rate and naturalness

Here we report AF analyses of fast versus slow speech, as well as natural versus synthesized speech. The analyses are designed to measure more stringent hypotheses about perceivably different effects of these manipulations on hierarchical temporal structure. Specifically, faster speech is predicted to result in less variability across all perceptible timescales, which should correspond with shallower, flatter AF functions. By contrast, synthesized speech is predicted to result in less variability in the longer timescales only, which should lead to shallower but more curved AF functions due to selective effects on longer timescales. The effect of speech rate is tested using both algorithmic and experimental manipulations, whereas the effect of naturalness is tested using two different algorithmic manipulations. For the latter, we compare results with synthesized versus sine wave speech [18]. Sine wave speech is a synthetic control that retains most of the hierarchical temporal structure in the original signal.

2. Methods

Analyses of speaking rate were based on Barack Obama's address at George Mason University on the 21st Century Economy (1/08/09, 17:08 mins). The *élastique* algorithm (https://products.zplane.de/) was used to manipulate speaking rate without affecting the vocal pitch. The "fast" condition was 2x faster than the original recording, and the "slow" condition was 2x slower. In addition to these algorithmic manipulations, an experiment was conducted in which ten University of California students read two excerpts from the speech off a teleprompter. Half of the participants read the first excerpt at a slow pace and the second at a fast pace, and vice versa for the other half. On average, the fast-paced and slow-paced excerpts took 4.5 and 10.1 min to complete, respectively. Participants were instructed to read the speech from the teleprompter as smoothly as possible, and their readings were recorded for subsequent acoustic analyses.

Analyses of naturalness were based on ten recordings of TED talks (mean length = 6.41 min, SD = 1.14 min) reported by Kello et al. [14]. The TED intro and outro theme was trimmed from the recordings, along with any applause at the beginnings or ends of the talks. A synthesized version of each talk was created by submitting the transcript to Google speech synthesis, and recording the output. The synthesized versions (mean length = 6.62 min,

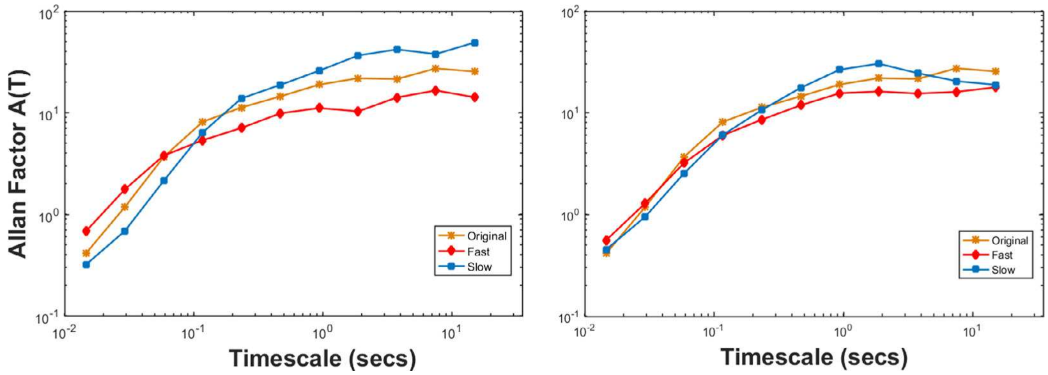

Fig. 1. Left: AF functions of the original Obama speech, and fast and slow versions. Right: AF functions of the fast and slow teleprompter conditions.

SD = 1.16 min) were recorded using GarageBand version 10.1.0. GarageBand was also used to set the lengths of the synthesized recordings roughly equal to the original recordings (within +/− 30 s). Lastly, sine wave speech recordings (mean length = 6.46 min, SD 1.16 min) were created from the ten trimmed TED talks by using the Matlab sine wave speech code provided by Ellis [7], with default parameters provided by Haskins Laboratories. The software tracks speech formants and assigns a single sine wave to each one. The sine wave amplitudes and frequencies are modulated to track the formants over time. The result is a combination of whistling sounds that preserve most of temporal structure in speech. Sine wave speech is typically perceived as speech-like, but the words spoken are difficult to discern unless the listener is given information about what is being said.

3. Results

Audio recordings were analyzed using the same method as reported in Kello et al. [14]. Details can be found there, but briefly: Each recording was divided into four-minute segments, and analyses were averaged across segments to yield a single AF function per recording. The Hilbert envelope was calculated for each segment and peaks above threshold were analyzed as time series of acoustic events. An AF function was computed for each segment:

$$A(T) = \frac{\left\langle (N_i(T) - N_{i+1}(T))^2 \right\rangle}{2\langle N_i(T)\rangle},$$

where T is the timescale, $N_i(T)$ is the event count in each window i, and $A(T)$ is AF variance. AF variance captures the degree of event clustering at a given timescale, and for time series with nested clustering, $A(T)$ increases with T. Self-similar clustering across timescales yields a power law, $A(T) \sim T^{\alpha}$, where $0 < \alpha < 2$. The AF function was computed for 11 values of T in between 15 ms and 15 s, logarithmically spaced to compute the orthonormal basis.

AF functions for speaking rate analyses are shown in Fig. 1. The left panel shows the effect of algorithmic speaking rate manipulations on the original Obama recording, and the right panel shows mean AF functions for the slow and fast teleprompter conditions, with the original Obama recording as a reference. AF variance for the Obama recording steadily increased as a function of timescale, consistent with analyses of TED talk recordings reported by Kello et al. [14]. Falk and Kello [8] found evidence to suggest that this AF shape is common to speech because it reflects the nesting of linguistic units like syllables in words, words in phrases, and phrases in sentences. Fig. 1 shows that an algorithmic increase in speaking rate causes clustering to generally shift left into the shorter timescales, whereas an algorithmic decrease causes a rightward shift into the longer timescales. Fig. 1 also shows that the teleprompter had a similar effect, except that there was a drop in AF variance at the longest timescales for slow speaking rates. We hypothesize that this drop comes from the artificially even pace of speaking caused by the slow, even pace of the teleprompter. This evenness creates isochrony and isochrony reduces clustering and hence AF variance. We leave it for future research to test this hypothesis explicitly.

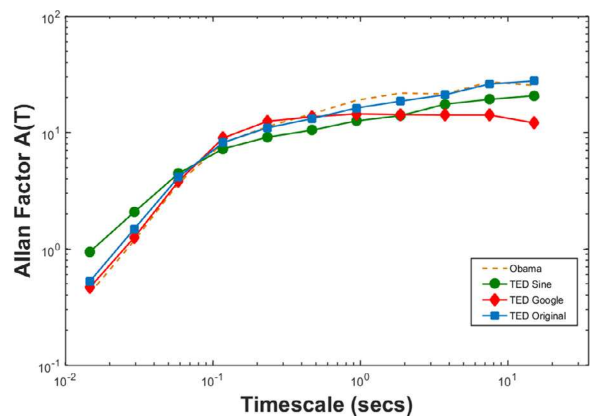

Fig. 2. Mean AF functions for TED talks and their two different synthesized versions, Google text-to-speech and sine wave speech. The AF function for Obama's speech is shown for comparison.

AF functions for naturalness analyses are shown in Fig. 2. The mean AF function for the original TED talk recordings has the same basic shape as that for the original Obama recording. This similarity is consistent with Kello et al. [14] who found that monologues have common, distinctive AF functions compared with dialogues and singing—TED talks and the Obama speech are both types of monologues. AF functions for synthesized versions of TED talks were very similar to the original recordings in the shorter timescales, but they diverged in the longer timescales. Specifically, synthesized AF functions were flat compared with original recordings, which indicates a lack of nested clustering in timescales corresponding with prosody and intonation. By contrast, AF functions for sine wave speech had the same overall shape as the TED talk recordings from which they were created, with a slight leftward shift of clustering as if the sine wave speech rate was faster than the original recording.

The perceptual distinction between natural and synthesized speech is very clear, as is the distinction between slow versus fast speaking rates. Moreover, these two dimensions of variation are perceptually distinct from each other. The effects of speaking rate and naturalness were also different from each other, as verbally described above, but it is necessary to quantify this difference to better understand it and relate it to complexity matching. To do

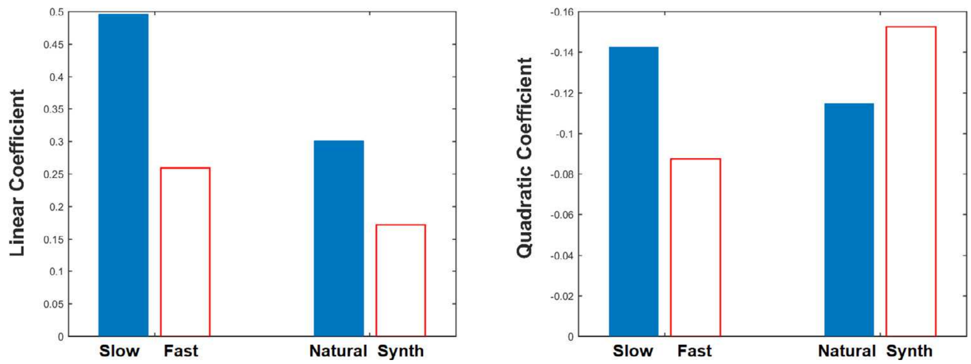

Fig. 3. Linear and quadratic coefficients for fast versus slow speech, and natural versus synthesized speech. The two different manipulations had the same effect on linear coefficients, but opposite effects on quadratic coefficients.

so, we fitted a second-order polynomial to each individual AF function, which allowed us to capture their convex shapes in terms of linear and quadratic coefficients.

Coefficients are plotted in Fig. 3 for fast and slow speaking rates, as well as natural and synthesized speech. The graph shows that speaking rate had the same effect on linear coefficients but opposite effects on quadratic coefficients. Fast speech was comparable to synthesized speech in that linear coefficients were closer to zero compared with slow speech and natural speech, respectively. This similar result was due to the overall flattening effect of these conditions. However, fast speech was less convex than slow speech, whereas synthesized speech was more convex than natural speech. This difference was due to the selective effect of synthesis on longer timescales, versus the overall effect of speaking rate across all measured timescales. Finally, sine wave synthesis had a small effect on coefficients akin to the effect of fast speech. It would be interesting to test whether sine wave is perceived as being faster than normal speech, even though the same signal variations unfold over the same time periods.

4. Discussion

In the present study, we investigated the effect of manipulating speaking rate and naturalness on hierarchical temporal structure in speech. Using AF analysis, we showed that nested clustering in peak amplitude events is affected differently by these two manipulations—changes in speaking rate shifts the entire measured hierarchy into shorter or longer timescales, whereas changes in naturalness flatten or steepen the longer timescales of the hierarchy, i.e. on the order of seconds and longer. Other studies have shown that acoustic events in speech appear to be crucial events [1], including a recent study by Pease et al. [17] in the present special issue edited by Grigolini [10]. Taken together, these studies suggest that neural and perceptual processes may be highly responsive to speech inputs by means of complexity matching. Specifically, power laws in neural and perceptual dynamics may take the general shape of power laws in speech dynamics by means of complexity matching, while having distinct trajectories because of myriad differences in neural versus acoustic "substrate", so to speak. The present results are consistent with this application of complexity matching, in that the different perceptual experiences associated with speaking rate and naturalness have corresponding differences in hierarchical temporal structure. These perceptual differences may have their roots in complexity matching of auditory networks with incoming speech signals.

The application of complexity matching to speech perception leads to questions about how power laws in auditory networks are affected when temporal structures in speech signals do not follow a single power law. Kello et al. [14] showed that many categories of speech and music deviate from power law AF functions. In fact, the only categories that closely followed a power law in nested event clustering were classical music and thunderstorms. Monologues like those analyzed herein were consistently found to have a distinct flattening in the longer timescales, and the shape of this deviation varies as a function of speaking rate and naturalness. What do such deviations imply for complexity matching?

One possibility is that neural and perceptual dynamics become less responsive to speech dynamics when they deviate from a power law, because brains are attuned to power laws in sensory inputs. Another possibility is that neural dynamics bend along with the dynamics of speech being listened to. The latter would correspond to a neural correlate of perceiving and following the sounds of speech. The same question can also be asked of music, with the same possible hypotheses [6]. Indeed, the effect of prosody on temporal hierarchies in speech has been argued to have an analog in music [11,16]. This analog leads to the idea that music perception, as hypothesized for speech perception, may be partly supported by a form of complexity matching that enables temporal hierarchies in neural dynamics to conform to those of speech and music.

Acknowledgments

This study stemmed from work presented at a conference supported by the Army Research Office through grant W911NF-16-1-0461. We thank the conference organizers and attendees for their fruitful discussions and interactions. The study was also supported by a National Science Foundation Research Training grant, award number 1633722.

References

[1] Abney DH, Paxton A, Dale R, Kello CT. Complexity matching in dyadic conversation. *J Exp Psychol* 2014;143(6):2304.
[2] Baronchelli A, Ferrer-i-Cancho R, Pastor-Satorras R, Chater N, Christiansen MH. Networks in cognitive science. *Trends Cognit Sci* 2013;17(7):348–60.
[3] Chen S, Hsieh C, Chiang C, Hsiao H, Wang Y, Liao Y, Yu H. Modeling of speaking rate influences on mandarin speech prosody and its application to speaking rate-controlled TTS. *IEEE/ACM Trans Audio Speech Lang Process (TASLP)* 2014;22(7):1158–71.
[4] Dellwo V, Wagner P. Relationships between rhythm and speech rate. In: International congress of phonetic sciences, Barcelona; 2003. *Journal of the International Phonetic Association. ICPhS-15*, p. 471–474.
[5] Ding N, Melloni L, Zhang H, Tian X, Poeppel D. Cortical tracking of hierarchical linguistic structures in connected speech. *Nat Neurosci* 2015;19:158. https://www.nature.com/articles/nn.4186#supplementary-information.
[6] Ding N, Patel AD, Chen L, Butler H, Luo C, Poeppel D. Temporal modulations in speech and music. *Neurosci Biobehav Rev* 2017.
[7] Ellis DPW, Sinewave Speech Analysis/Synthesis in Matlab. 2004. Web resource, available: http://www.ee.columbia.edu/ln/labrosa/matlab/sws/.
[8] Falk S, Kello CT. Hierarchical organization in the temporal structure of infant-direct speech and song. Cognition 2017;163:80–6.
[9] Gibson JJ. The ecological approach to visual perception. Boston: Houghton Mifflin; 1979.

[10] Grigolini P. Call for papers: Special issue on evolutionary game theory of small groups and their larger societies. *Chaos Solitons Fractals* 2017;103:371–3.

[11] Hausen M, Torppa R, Salmela VR, Vainio M, Srkm T. Music and speech prosody: a common rhythm. *Front Psychol* 2013;4:566.

[12] Jun S. The effect of phrase length and speech rate on prosodic phrasing. In: Paper presented at the *proceedings of the XVth international congress of phonetic sciences*; 2003. p. 483–6.

[13] Kello CT, Brown GD, Ferrer-i-Cancho R, Holden JG, Linkenkaer-Hansen K, Rhodes T, Van Orden GC. Scaling laws in cognitive sciences. *Trends Cognit Sci* 2010;14(5):223–32.

[14] Kello CT, Dalla Bella S, Médé B, Balasubramaniam R. Hierarchical temporal structure in music, speech and animal vocalizations: Jazz is like a conversation, humpbacks sing like hermit thrushes. *J R Soc Interface* 2017;14(135):20170231.

[15] Marmelat V, Delignires D. Strong anticipation: complexity matching in interpersonal coordination. *Exp Brain Res* 2012;222(1-2):137–48.

[16] Palmer C, Hutchins S. What is musical prosody? *Psychol Learn Motivation* 2006;46:245–78.

[17] Pease A, Mahmoodi K, West BJ. Complexity measures of music. *Chaos Solitons Fractals* 2018;108:82–6.

[18] Remez RE, Rubin PE, Pisoni DB, Carrell TD. Speech perception without traditional speech cues. Science 1981;212(4497):947–9.

[19] Stephen DG, Stepp N, Dixon JA, Turvey MT. Strong anticipation: Sensitivity to long-range correlations in synchronization behavior. *Phys A Stat Mech Appl* 2008;387(21):5271–8.

[20] Tseng C, Lee Y. Speech rate and prosody units: evidence of interaction from mandarin Chinese. Speech prosody 2004, international conference Paper presented at the; 2004.

[21] West BJ, Geneston EL, Grigolini P. Maximizing information exchange between complex networks. *Phys Rep* 2008;468(1):1–99.

[22] West B, Bologna M, Grigolini P. Physics of fractal operators. Springer Science & Business Media; 2012.

[23] Ze H, Senior A, Schuster M. Statistical parametric speech synthesis using deep neural networks. In: Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on; 2013. p. 7962–6.