Considerations for Child Speech Synthesis for Dialogue Systems

Kallirroi Georgila

Abstract We present a number of important issues for consideration with regard to child speech synthesis for dialogue systems. We specifically discuss challenges in building child synthetic voices compared to adult synthetic voices, synthesizing expressive conversational speech, and evaluating speech synthesis quality.

1 Introduction

Although children are an important user group for dialogue system applications, there has been relatively little work on building artificial agents designed specifically for interacting with children [38, 34, 49, 12, 22, 4, 6], compared to the vast amount of effort put into building dialogue systems for adults. This needs to change in order to make dialogue systems more inclusive and accessible. In this position/survey paper our focus is on one aspect related to developing dialogue systems, namely, child speech synthesis, i.e., building voices that sound like children [53, 23, 46].

Speech synthesis (also known as text-to-speech synthesis) is the automatic process of converting natural language text into speech [45]. Speech synthesis has many potential applications [15]. Here we are specifically concerned with speech synthesis as a means for providing spoken dialogue systems, virtual humans, and robots with child-like synthetic voices. This is because we may want to have dialogue systems such as educational games where children interact with peers, i.e., child-like artificial agents as well as other children [4].

Below we discuss challenges in building child synthetic voices compared to adult synthetic voices, synthesizing expressive conversational speech, and evaluating speech synthesis quality.

Kallirroi Georgila, USC Institute for Creative Technologies, 12015 Waterfront Drive, Los Angeles, CA 90094 USA, e-mail: kgeorgila@ict.usc.edu

2 Challenges in Building Child Synthetic Voices

Child speech synthesis has generally been studied less than child speech recognition [44, 43] or paralinguistic analysis of child speech [41, 32], but they are all more challenging than their adult counterparts. This is because of the larger intra- and inter-speaker variability, with regard to acoustic and linguistic characteristics, in comparison with adult speech. Furthermore, there is the additional challenge of lack of adequate and appropriate child speech resources [43, 42].

We can also draw an analogy between children's and older adults' voices in the sense that they both pose challenges compared to the general population. For older adults, chronological age is a relatively poor predictor of anatomical, physiological, and cognitive changes [48, 55, 56, 18, 20, 19, 21]. For children, chronological age can be a better predictor of the above changes, but similar to older adults, there is large variability. For example, age can have anatomical impacts. Shorter vocal tracts and smaller vocal folds in children result in higher fundamental and formant frequencies than for adults [40, 44], which in turn affects how a child voice sounds, hence the use of vocal tract length normalization for child speech processing [40, 44]. Furthermore, children tend to speak more slowly than adults, and there is larger variability in their speaking rates, because their articulators are not fully developed yet [39]. Children may also use more imaginative words and ungrammatical phrases [24], and children's speech is characterized by larger variability with respect to speech disfluencies including hesitations, repetitions, and revisions [39, 53, 43].

Over the years various techniques have been used for speech synthesis. The most popular recent methods are data-driven such as unit selection [8, 33, 51], Hidden Markov Model (HMM)-based speech synthesis [62, 63], and more recently deep learning-based speech synthesis [47, 52, 35]. All the above speech synthesis methods are data-hungry and, to a greater or lesser extent, require noise-free recordings of phonetically balanced and consistently read speech.

Watts et al. [53] list several problems typical in collecting data for child speech synthesis: (1) Getting a child to a studio for recordings is more difficult than an adult voice talent. Consequently, recordings are done at home with considerable background noise. (2) Consistently recorded speech requires a certain level of vocal and emotional control that children do not have. It is very hard to convince children to record long sessions, which results in a higher number of short recording sessions, and consequently larger inter-session variability. (3) To create synthetic voices that can reliably generate all the phonemes of a language in different contexts, phonetically balanced data are necessary. Thus voice talents have to record a large number of prompts to ensure high phonetic coverage. The problem is that this kind of texts is very different from the type of texts that children are used to reading, such as stories and fairy tales. However, stories and fairy tales are not phonetically balanced, thus if they are used for data collection, this will result in poor phonetic coverage in the recorded data.

Similarly, Govender et al. [23] note that it is a major challenge to find children willing to record hours of speech, and even when a suitable candidate is found, the resulting recordings are not adequate or are lacking in terms of quality.

It is mentioned on the website of Acapela[1], a company specializing in speech synthesis (including child speech synthesis), that it can take months to identify the right voice talents for building child synthetic voices. Once such a child voice talent has been identified, the usual process used for building adult professional voices has to be adapted to the child's behaviour and habits.

A natural question that arises is: how can we minimize our reliance on high-quality child speech data by leveraging other sources of audio such as adult speech data?

One way to leverage adult speech data is by using speaker-adaptive HMM-based speech synthesis [57, 58] where the process is as follows: First an average-voice model is built using speech from multiple speakers, or a background model is built from one speaker. Then, using small amounts of data from the target speaker (in our case a child), we can adapt the parameters of the average-voice model or the background model, to capture the voice characteristics of this target speaker. Speaker-adaptive HMM-based speech synthesis has been used for both child speech synthesis [53, 23] and child speech recognition [26].

Watts et al. [53] found that child speaker-dependent voices performed worse than adult average-voice models adapted to the child target speaker data. This finding agreed with previous work with adult target speakers. However, in the case of child target speakers, more target speaker data were needed to achieve reasonable similarity to the child target speaker, which suggests that we need better average-voice models for child speech.

Deep learning has also been used for child speech synthesis either as a means to generate child synthetic voices [29, 60], or in order to augment real child data with synthetic child speech data to improve child speech recognition performance [27].

Note that using adult speech data to help with child speech synthesis may produce results that do not agree with our intuitive assumptions. For example, Govender et al. [23] found that using a gender-independent average-voice model resulted in better child speech synthesis than gender-dependent (either male or female) average-voice models. However, one would intuitively expect that a female average-voice would be a better choice given that the fundamental frequency of children's voices is closer to adult females than adult males.

Overall, this is an active area of research and there is no consensus about the best approach to building child synthetic voices. It is not even clear if deep learning approaches are superior to HMM-based models in all cases (given that deep learning methods require much more data than HMM-based approaches, which is a major issue for child speech synthesis), or what kind of adult speech data should be used to augment child speech data.

Several companies offer commercial child synthetic voices, but how to efficiently build child synthetic voices, without requiring expensive and time-consuming data collection, is still an open research problem, as well as how to deal with different accents and languages (including under-resourced languages).

[1] https://www.acapela-group.com/voices/children-voices/

3 Synthesizing Expressive Conversational Speech

Current state-of-the-art speech synthesizers can generate high-quality synthetic speech that sounds like reading from text in terms of naturalness and intelligibility [31], but are not that good at synthesizing expressive conversational speech. Spontaneous conversational speech exhibits characteristics that are very hard to model in speech synthesis, e.g., pronunciation variation [54], speech disfluencies (repetitions, repairs, hesitations) [10, 14, 17, 61], paralinguistics (laughter, breathing) [9, 41], etc.

Some previous work on conversational speech synthesis has focused on filled pauses (e.g., "uh", "um"), in particular, on predicting where to insert filled pauses in an utterance so that it sounds natural [2], how to synthesize such filled pauses [1, 2, 3], and how to model sequences of pronunciation variants to generate a more conversational style of speech [54].

The Google Duplex demo[2] exhibited impressive capabilities in generating conversational synthetic speech, especially fillers, based on WaveNets [47], but it is not clear how exactly these fillers were modelled.

Nevertheless the state-of-the-art is still far from human-like conversational speech with hesitations, revisions, restarts, and repetitions, and thus deep learning-based conversational speech synthesis is an active area of research [25, 28].

As mentioned above, children's speech is characterized by large variability with respect to speech disfluencies including hesitations, repetitions, and revisions [39, 53, 43]. Thus in order to build realistic child synthetic voices we need to make more progress towards synthesizing conversational speech phenomena.

Again, on the website of Acapela, it is mentioned that children's typical exclamations and sounds are recorded to create a natural and spontaneous audio result, but without providing further details. Thus it is not clear if these recordings are used for training spontaneous speech models via machine learning, or are just used as canned audio.

Similar to conversational speech, emotional speech is another area where synthetic speech is lacking in quality [5]. Most research on emotional speech synthesis uses data that contain acted emotions, i.e., actors are asked to simulate emotions such as happiness, sadness, anger, etc. However, such simulated emotions differ significantly from emotions experienced in the real world [13]. Due to ethical and privacy concerns, a major challenge in emotional child speech synthesis (and emotional speech synthesis in general) is acquiring speech that exhibits real emotions.

Overall, more emphasis should be placed on improving the state-of-the-art on expressive conversational speech synthesis for both adults and children. This is necessary for building artificial agents that can realistically simulate children and engage in conversation with real children. But, as discussed below in section 5, we also need to be prepared to deal with ethical challenges, and ensure that this technology is not misused to help criminals pose as children.

[2] https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

4 Speech Synthesis Quality Evaluation Considerations

The current practice in speech synthesis evaluation is to ask human raters to rate isolated audio clips, usually in terms of naturalness and intelligibility [31], likability [16], or how conversational it sounds [2, 3, 16], without extended exposure to a voice. This approach can certainly inform us about the general quality of a synthetic voice; but it cannot necessarily provide any insight about the appropriateness of this voice for a task that requires that the listener be exposed to that voice for a considerable amount of time. Furthermore, as the environments where dialogue systems are deployed become increasingly immersive involving multiple agents, it becomes critical to determine how subjective perceptions of a voice change, if voice exposure is sporadic and interleaved with other voices.

To that end Pincus et al. [37] found that synthetic voices' likability and naturalness perceptions degrade based on time/continuity of exposure, while human voices' likability and naturalness perceptions improve with increasing time/continuity. Betz et al. [7] showed that due to its conversational nature, hesitation synthesis needs interactive evaluation (similar to [37]). Furthermore, their results suggest that synthetic hesitations can improve task performance, but to avoid likability issues, an elaborate hesitation strategy is necessary.

These studies and other related research [50] show that we need to revise our current speech synthesis evaluation practices. It is not clear how tolerant of poor-quality synthetic speech children are, especially for extended exposure. It is also application-dependent whether an artificial agent interacting with children should be equipped with an adult or child synthetic voice.

5 Conclusion

We discussed challenges in building child synthetic voices compared to adult synthetic voices, synthesizing expressive conversational speech, and evaluating speech synthesis quality. Apart from these technological challenges, there are also ethical concerns and challenges, especially when such technology is targeted at children and uses speech recordings from children. Although researchers work on detecting synthetic voice spoofing [30, 64] and audio deepfakes [36] for adults, the troubling danger of criminals using speech synthesis to pose as children is currently not being adequately addressed. Despite the challenges of creating high-quality child synthetic voices and detecting deceptive use of such voices, there are many positive applications, including synthetic voices for children with disabilities [59, 11], and intelligent tutoring systems that play the role of a fellow student [4]. Therefore, it is important that speech and dialogue researchers place more emphasis on these areas.

Acknowledgements This work was partly supported by the U.S. Army. Statements and opinions expressed and content included do not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

1. Adell, J., Bonafonte, A., Escudero, D.: Disfluent speech analysis and synthesis: A preliminary approach. In: Proceedings of the International Conference on Speech Prosody. Dresden, Germany (2006)
2. Andersson, S., Georgila, K., Traum, D., Aylett, M., Clark, R.A.J.: Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In: Proceedings of the International Conference on Speech Prosody. Chicago, Illinois, USA (2010)
3. Andersson, S., Yamagishi, J., Clark, R.A.J.: Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis. Speech Communication **54**(2), 175–188 (2012)
4. Aslan, S., Agrawal, A., Alyuz, N., Chierichetti, R., Durham, L.M., Manuvinakurike, R., Okur, E., Sahay, S., Sharma, S., Sherry, J., Raffa, G., Nachman, L.: Exploring Kid Space in the wild: a preliminary study of multimodal and immersive collaborative play-based learning experiences. Educational technology research and development **70**, 205–230 (2022)
5. Barra-Chicote, R., Yamagishi, J., King, S., Montero, J.M., Macias-Guarasa, J.: Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Communication **52**(5), 394–404 (2010)
6. Beelen, T., Velner, E., Ordelman, R., Truong, K.P., Evers, V., Huibers, T.: Designing conversational robots with children during the pandemic. In: 6th International and Interdisciplinary Perspectives on Children Recommender and Information Retrieval Systems (KidRec), Information Retrieval Systems for Children in the COVID-19 Era; co-located with ACM IDC. Braga, Portugal (2022)
7. Betz, S., Carlmeyer, B., Wagner, P., Wrede, B.: Interactive hesitation synthesis: Modelling and evaluation. Multimodal Technologies and Interaction **2**(1), 9 (2018)
8. Black, A.W., Taylor, P.: Automatically clustering similar units for unit selection in speech synthesis. In: Proceedings of Eurospeech, pp. 601–604. Rhodes, Greece (1997)
9. Campbell, N.: Conversational speech synthesis and the need for some laughter. IEEE Transactions on Audio, Speech, and Language Processing **14**(4), 1171–1178 (2006)
10. Core, M.G., Schubert, L.K.: A syntactic framework for speech repairs and other disruptions. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), pp. 413–420. College Park, Maryland, USA (1999)
11. Creer, S., Cunningham, S., Green, P., Yamagishi, J.: Building personalised synthetic voices for individuals with severe speech impairment. Computer Speech and Language **27**(6), 1178–1193 (2013)
12. Davison, D.P., Wijnen, F.M., Charisi, V., van der Meij, J., Evers, V., Reidsma, D.: Working with a social robot in school: A long-term real-world unsupervised deployment. In: Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI), pp. 63–72. Cambridge, UK (2020)
13. Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P.: Emotional speech: Towards a new generation of databases. Speech Communication **40**(1-2), 33–60 (2003)
14. Georgila, K.: Using integer linear programming for detecting speech disfluencies. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Companion Volume: Short Papers, pp. 109–112. Boulder, Colorado, USA (2009)
15. Georgila, K.: Speech synthesis: State-of-the-art and challenges for the future. In: J.K. Burgoon, N. Magnenat-Thalmann, M. Pantic, A. Vinciarelli (eds.) Social Signal Processing, pp. 257–272. Cambridge University Press (2017)
16. Georgila, K., Black, A.W., Sagae, K., Traum, D.: Practical evaluation of human and synthesized speech for virtual human dialogue systems. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 3519–3526. Istanbul, Turkey (2012)
17. Georgila, K., Wang, N., Gratch, J.: Cross-domain speech disfluency detection. In: Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), pp. 237–240. Tokyo, Japan (2010)

18. Georgila, K., Wolters, M., Karaiskos, V., Kronenthal, M., Logie, R., Mayo, N., Moore, J.D., Watson, M.: A fully annotated corpus for studying the effect of cognitive ageing on users' interactions with spoken dialogue systems. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 938–944. Marrakech, Morocco (2008)
19. Georgila, K., Wolters, M., Moore, J.D., Logie, R.H.: The MATCH corpus: A corpus of older and younger users' interactions with spoken dialogue systems. Language Resources and Evaluation **44**(3), 221–261 (2010)
20. Georgila, K., Wolters, M.K., Moore, J.D.: Simulating the behaviour of older versus younger users when interacting with spoken dialogue systems. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics - Human Language Technologies (ACL-HLT), Short Papers, pp. 49–52. Columbus, Ohio, USA (2008)
21. Georgila, K., Wolters, M.K., Moore, J.D.: Learning dialogue strategies from older and younger simulated users. In: Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), pp. 103–106. Tokyo, Japan (2010)
22. Gillet, S., van den Bos, W., Leite, I.: A social robot mediator to foster collaboration and inclusion among children. In: Proceedings of Robotics: Science and Systems. Corvalis, Oregon, USA (2020)
23. Govender, A., de Wet, F., Tapamo, J.R.: HMM adaptation for child speech synthesis. In: Proceedings of Interspeech, pp. 1640–1644. Dresden, Germany (2015)
24. Gray, S.S., Willett, D., Lu, J., Pinto, J., Maergner, P., Bodenstab, N.: Child automatic speech recognition for US English: Child interaction with living-room-electronic-devices. In: Proceedings of the Workshop on Child Computer Interaction (WOCCI). Singapore (2014)
25. Guo, H., Zhang, S., Soong, F.K., He, L., Xie, L.: Conversational end-to-end TTS for voice agents. In: Proceedings of the IEEE Spoken Language Technology Workshop (SLT), pp. 403–409 (2021)
26. Hagen, A., Pellom, B., Hacioglu, K.: Generating synthetic children's acoustic models from adult models. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), Companion Volume: Short Papers, pp. 77–80. Boulder, Colorado, USA (2009)
27. Hasija, T., Kadyan, V., Guleria, K.: Out domain data augmentation on Punjabi children speech recognition using Tacotron. In: Proceedings of the International Conference on Mathematics and Artificial Intelligence (ICMAI). Chengdu, China (2021)
28. Hu, Y., Liu, R., Gao, G., Li, H.: FCTalker: Fine and coarse grained context modeling for expressive conversational speech synthesis. In: arXiv:2210.15360 (2022)
29. Jia, N., Zheng, C., Sun, W.: Speech synthesis of children's reading based on CycleGAN model. In: Proceedings of the the International Symposium on Electronic Information Technology and Communication Engineering (ISEITCE). Jinan, China (2020)
30. Kamble, M.R., Sailor, H.B., Patil, H.A., Li, H.: Advances in anti-spoofing: from the perspective of ASVspoof challenges. APSIPA Transactions on Signal and Information Processing **9** (2020)
31. Karaiskos, V., King, S., Clark, R.A.J., Mayo, C.: The Blizzard challenge 2008. In: Proceedings of the Blizzard Challenge Workshop. Brisbane, Australia (2008)
32. Kaya, H., Verkholyak, O., Markitantov, M., , Karpov, A.: Combining clustering and functionals based acoustic feature representations for classification of baby sounds. In: Proceedings of the ACM Workshop on Bridging Social Sciences and AI for Understanding Child Behavior - included in the Companion Proceedings of the ACM International Conference on Multimodal Interaction (ICMI), pp. 509–513. Online (2020)
33. Kishore, S.P., Black, A.W.: Unit size in unit selection speech synthesis. In: Proceedings of Interspeech, pp. 1317–1320. Geneva, Switzerland (2003)
34. Leite, I., Pereira, A., Lehman, J.F.: Persistent memory in repeated child-robot conversations. In: Proceedings of the Conference on Interaction Design and Children (IDC), pp. 238–247. Stanford, California, USA (2017)
35. Li, H., Yamagishi, J.: Multi-metric optimization using generative adversarial networks for near-end speech intelligibility enhancement. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3000–3011 (2021)

36. Pan, J., Nie, S., Zhang, H., He, S., Zhang, K., Liang, S., Zhang, X., Tao, J.: Speaker recognition-assisted robust audio deepfake detection. In: Proceedings of Interspeech, pp. 4202–4206. Incheon, Korea (2022)
37. Pincus, E., Georgila, K., Traum, D.: Which synthetic voice should I choose for an evocative task? In: Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL), pp. 105–113. Prague, Czech Republic (2015)
38. Potamianos, A., Narayanan, S.: Spoken dialogue systems for children. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 197–200. Seattle, Washington, USA (1998)
39. Potamianos, A., Narayanan, S., Lee, S.: Automatic speech recognition for children. In: Proceedings of Eurospeech, pp. 2371–2374. Rhodes, Greece (1997)
40. Qian, Y., Wang, X., Evanini, K., Suendermann-Oeft, D.: Improving DNN-based automatic recognition of non-native children's speech with adult speech. In: Proceedings of the Workshop on Child Computer Interaction (WOCCI). San Francisco, California, USA (2016)
41. Rao, H., Clements, M.A., Li, Y., Swanson, M.R., Piven, J., Messinger, D.S.: Paralinguistic analysis of children's speech in natural environments. In: J. Rehg, S. Murphy, S. Kumar (eds.) Mobile Health, pp. 219–238. Springer (2017)
42. Rumberg, L., Gebauer, C., Ehlert, H., Wallbaum, M., Bornholt, L., Ostermann, J., Lüdtke, U.: kidsTALC: A corpus of 3- to 11-year-old German children's connected natural speech. In: Proceedings of Interspeech, pp. 5160–5164. Incheon, Korea (2022)
43. Shivakumar, P.G., Narayanan, S.: End-to-end neural systems for automatic children speech recognition: An empirical study. Computer Speech and Language **72** (2022)
44. Shivakumar, P.G., Potamianos, A., Lee, S., Narayanan, S.: Improving speech recognition for children using acoustic adaptation and pronunciation modeling. In: Proceedings of the Workshop on Child Computer Interaction (WOCCI). Singapore (2014)
45. Taylor, P.: Text-to-speech synthesis. Cambridge University Press (2009)
46. Terblanche, C., Harty, M., Pascoe, M., Tucker, B.V.: A situational analysis of current speech-synthesis systems for child voices: A scoping review of qualitative and quantitative evidence. Applied Sciences, Special Issue on Applications of Speech and Language Technologies in Healthcare **12**(11), 5623 (2022)
47. van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K.: WaveNet: A generative model for raw audio. In: Proceedings of the 9th ISCA Speech Synthesis Workshop (SSW 9) (2016)
48. Vipperla, R., Wolters, M., Georgila, K., Renals, S.: Speech input from older users in smart environments: Challenges and perspectives. In: Proceedings of Universal Access in Human-Computer Interaction, HCI International, Lecture Notes in Computer Science, Vol. 5615, pp. 117–126. Springer Berlin Heidelberg (2009)
49. Vollmer, A.L., Read, R., Trippas, D., Belpaeme, T.: Children conform, adults resist: a robot group induced peer pressure on normative social conformity. Science Robotics **3**(21) (2018)
50. Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Éva Székely, Tånnander, C., Voße, J.: Speech synthesis evaluation — State-of-the-art assessment and suggestion for a novel research program. In: Proceedings of the 10th ISCA Speech Synthesis Workshop (SSW 10), pp. 105–110 (2019)
51. Wang, W.Y., Georgila, K.: Automatic detection of unnatural word-level segments in unit-selection speech synthesis. In: Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 289–294. Big Island, Hawaii, USA (2011)
52. Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., Saurous, R.A.: Tacotron: Towards end-to-end speech synthesis. In: Proceedings of Interspeech, pp. 4006–4010. Stockholm, Sweden (2017)
53. Watts, O., Yamagishi, J., King, S., Berkling, K.: Synthesis of child speech with HMM adaptation and voice conversion. IEEE Transactions on Audio, Speech, and Language Processing **18**(5), 1005–1016 (2010)

54. Werner, S., Hoffmann, R.: Spontaneous speech synthesis by pronunciation variant selection - A comparison to natural speech. In: Proceedings of Interspeech, pp. 1781–1784. Antwerp, Belgium (2007)
55. Wolters, M., Georgila, K., Moore, J.D., Logie, R.H., MacPherson, S.E., Watson, M.: Reducing working memory load in spoken dialogue systems. Interacting with Computers **21**(4), 276–287 (2009)
56. Wolters, M., Georgila, K., Moore, J.D., MacPherson, S.E.: Being old doesn't mean acting old: How older users interact with spoken dialog systems. ACM Transactions on Accessible Computing (TACCESS) **1**, Article No. 2 (2009)
57. Yamagishi, J., Nose, T., Zen, H., Ling, Z.H., Toda, T., Tokuda, K., King, S., Renals, S.: Robust speaker-adaptive HMM-based text-to-speech synthesis. IEEE Transactions on Audio, Speech, and Language Processing **17**(6), 1208–1230 (2009)
58. Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Guan, Y., Hu, R., Oura, K., Wu, Y.J., Tokuda, K., Karhila, R., Kurimo, M.: Thousands of voices for HMM-based speech synthesis-analysis and application of TTS systems built on various ASR corpora. IEEE Transactions on Audio, Speech, and Language Processing **18**(5), 984–1004 (2010)
59. Yamagishi, J., Veaux, C., King, S., Renals, S.: Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. Acoustical Science Technology **33**(1), 1–5 (2012)
60. Yoshida, S., Furuya, K., Mizuno, H.: Introducing speaker vectors for child speech synthesis in neural vocoders. In: Proceedings of CISIS: Complex, Intelligent and Software Intensive Systems, Lecture Notes in Networks and Systems, Vol. 497, pp. 538–547. Springer (2022)
61. Zayats, V., Ostendorf, M.: Giving attention to the unexpected: Using prosody innovations in disfluency detection. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pp. 86–95. Minneapolis, Minnesota, USA (2019)
62. Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A.W., Tokuda, K.: The HMM-based speech synthesis system (HTS) version 2.0. In: Proceedings of the ISCA Workshop on Speech Synthesis, pp. 294–299. Bonn, Germany (2007)
63. Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication **51**(11), 1039–1064 (2009)
64. Zhang, Y., Jiang, F., Duan, Z.: One-class learning towards synthetic voice spoofing detection. IEEE Signal Processing Letters **28**, 937–941 (2021)