INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Using speech synthesis to explain automatic speaker recognition: a new application of synthetic speech

Georgina Brown[1,2], *Christin Kirchhübel*[2] *and Ramiz Cuthbert*[1]

[1]Department of Linguistics and English Language, Lancaster University, Lancaster, UK
[2]Soundscape Voice Evidence, Lancaster, UK
g.brown5@lancaster.ac.uk, ck@soundscapevoice.com, r.cuthbert@lancaster.ac.uk

Abstract

Some speech synthesis systems make use of zero-shot adaptation to generate speech based on a target speaker. These systems produce speaker embeddings in the same way that speaker embeddings (often called 'x-vectors') are produced in automatic speaker recognition systems. This commonality between the two technologies could lower barriers that constrain the use of automatic speaker recognition systems in forensic speech analysis casework. A key barrier to the use of automatic speaker recognition in the forensic context is the issue of explainability, including what information about the voice a system uses in order to arrive at conclusions. This paper sets out a new approach that could be used to effectively communicate this type of information to audiences in the legal setting. Specifically, it is proposed that exposing listeners to synthetic speech produced by a zero-shot adaptation system could illustrate what aspects of the voice an automatic speaker recognition system captures.

Index Terms: speech synthesis, automatic speaker recognition, forensic speech science, explainability.

1. Introduction

Speech synthesis technology has moved away from concatenative and statistical parametric methods and has moved towards neural network-based methods. These neural network-based methods can generate speech samples that are generally considered to be of a higher quality [1]. The performance of these systems has traditionally been evaluated using Mean Opinion Scores (MOS) elicited from pools of human listeners [2]. In addition to gathering these types of Likert-style ratings, researchers are exploring the use of automatic evaluations of synthetic speech samples, and specific challenges have even been set up to innovate in this area (namely, the VoiceMOS Challenge 2022 [3]).

Speech synthesis has a wide range of applications including, for example, as an assistive technology tool for people with visual impairments, as an accessibility tool for telephone / banking services, or as a tool to enable voice reconstruction for people with medical conditions affecting their use of their voice. In view of these applications, the purpose of evaluating synthetic speech – be it by human listeners or automatic systems – is linked to the 'useability' of the technology. In this paper, we identify another purpose to evaluating synthetic speech, one which is not directly linked to the 'useability' of technology but rather one which feeds into the 'explainability' of technology. Speech synthesis systems belonging to a category called *zero-shot adaptation* [1] share a modelling component (speaker embeddings) with automatic speaker recognition systems. The outputs of such speech synthesis systems therefore reflect how automatic speaker recognition systems form models of speakers and provide indications of what information about voices is captured by an automatic speaker recognition system. Asking listeners to evaluate these synthetic samples then could be a tool to explaining to those listeners what automatic speaker recognition systems do.

Our proposal is motivated by the use of automatic speaker recognition systems in the legal setting and therefore is of particular interest to the forensic speech science community. Attempts to introduce results produced by automatic speaker recognition systems as evidence in criminal court proceedings in England and Wales has resulted in questions around the explainability of such systems. In particular, the court has queried how the functioning of these systems can be effectively communicated to a jury, which typically comprises 12 laymen and women. We consider that speech synthesis technology and, in particular, the evaluation of synthetic speech, can assist the forensic speech science community in addressing this question. In turn, it could bring down a perceived barrier around the use of automatic speaker recognition results as evidence in UK jurisdictions.

This paper aims to triangulate between three sub-areas: speech synthesis, automatic speaker recognition and forensic speech science, because we consider that one of these sub-areas (speech synthesis) could substantially benefit another (forensic speech science).

2. Background

This section describes how the zero-shot adaptation approach to speech synthesis could contribute to the explainability of automatic speaker recognition systems. Sometimes the term 'explainability' is separated from the term 'interpretability', and sometimes 'explainability' encapsulates 'interpretability'. For the purpose of the idea put forward in this paper, there is no need to separate these two concepts, and therefore we have opted to use one term, 'explainability', throughout.

We first discuss the emergence of speaker embeddings in zero-shot adaptation systems before discussing how outputs from these systems have been evaluated. The section then sets out the proposed new forensic application for these synthetic speech samples, i.e., to add to our toolkit for explaining automatic speaker recognition systems.

10.21437/Interspeech.2023-1013

2.1. Speaker embeddings in zero-shot adaptation speech synthesis

2.1.1. Their role within speech synthesis

Applications such as voice reconstruction and personalised speech-to-speech translation place more importance on synthesising the voice of a *particular* speaker (in contrast to speech synthesis systems used as accessibility tools for telephone services, for example). To create speaker-specific synthetic speech, many minutes of speech per speaker were previously required. Therefore, a key purpose of [4] was to find a way to reduce the amount of speech data needed from a target speaker. [4] were successful in this respect. Their approach of decoupling speaker modelling from the speech synthesis system reduced the amount of speech needed to just seconds. To do this, they took inspiration from the way speaker modelling is typically carried out in automatic speaker recognition systems by training a neural network on a speaker verification task. That trained neural network (or *speaker encoder network* is the term used in [4]) can then be used to produce fixed-dimensional embeddings (often called 'x-vectors' in the speaker recognition literature [5]). The intention behind the embedding is to capture the speaker-specific characteristics of the speech sample. This speaker model is then combined with the output of the speech synthesiser to produce speech that sounds like that speaker.

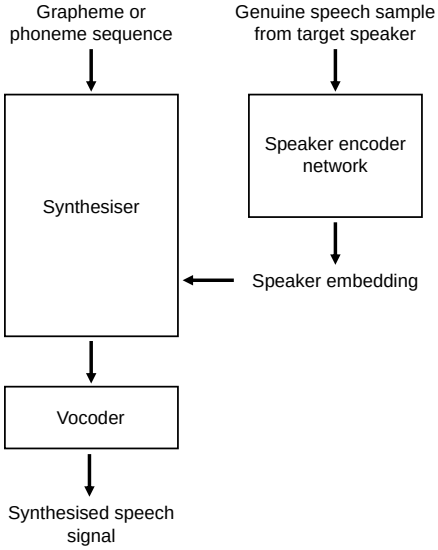

Figure 1: *Flow diagram of a zero-shot adaptation system (based on that presented in [4]).*

This method, now referred to as zero-shot adaptation, has been used in other works (e.g., [6]). Given that this method employs the same speaker modelling technique used in automatic speaker recognition systems, the output of zero-shot adaptation speech synthesis could demonstrate what, and how much, information about a speaker's voice is typically captured by automatic speaker recognition systems.

2.1.2. Evaluation of performance

In their own experiments, [4] used two ways of evaluating the similarity between genuine speech samples of a target speaker and synthesised samples for that speaker. On the one hand, they collected human evaluations (through MOS). For this, they used speech samples from the Voice Cloning Toolkit (VCTK) corpus [7]. They recorded an overall similarity MOS of 3.28 which fell between "moderately similar" and "very similar" on the scale. [4] concede that the possible degree of speaker similarity obtained from the human listeners may have been constrained because the speaker encoder was trained on North American English, while VCTK contains speakers of a range of English varieties (mostly British English varieties). On the other hand, [4] ran experiments involving an automatic speaker verification system. They reported an Equal Error Rate (EER) of 2.86% on a task involving real and synthetic speech of 10 speakers from the LibriSpeech database [8]. Based on these human and automatic evaluations, [4] concluded that "*the proposed model can generate speech that resembles the target speaker, but not well enough to be confusable with a real speaker*" [p. 7]. In view of evaluations carried out by others - see below - this conclusion may be regarded as rather conservative.

Speech synthesised by [4]'s system has been evaluated more widely through the ASVspoof 2019 and 2021 challenges [9, 10]. These ASVspoof challenges gather speech samples produced by a range of different "spoofing" methods (including speech synthesis systems) into a single dataset and then invite researchers to carry out automatic speaker recognition and spoofing detection experiments on that dataset. In works related to these ASVspoof challenges, the speech samples produced by [4]'s system have been highlighted as particularly problematic (when comparing them against other "spoofing" methods). [9] reported that [4]'s system yielded 57.73% EER in automatic speaker verification experiments. Further, it fell within the three most undetectable spoofing methods included in that challenge. This confirms that synthetic speech produced by [4]'s system both "tricks" automatic speaker recognition systems and is difficult to detect by spoofing detection systems.

The synthetic speech produced by [4]'s system has also proved difficult for human evaluators to detect. [11] gathered results from 165 listeners which established that this system was repeatedly producing speech samples that were mistaken for genuine human speech. [12] presented speech samples from spoofing methods included in the ASVspoof 2019 challenge dataset to an expert forensic phonetician for evaluation. [4]'s system was singled out as the spoofing method that produced the most human-like speech samples to the extent that they were indistinguishable from genuine human speech samples. 74% of these synthetic speech samples presented to the phonetician could not be confidently labelled as "spoofs", while 26% of these synthetic speech samples were evaluated as genuine human samples.

From [4]'s experiments and work on the ASVspoof datasets, it is clear that a zero-shot adaptation speech synthesis system generates speech that is both human-like and speaker-like. Based on its ability to generate speaker-specific synthetic speech, we propose that it can add a crucial contribution to explaining how automatic speaker recognition systems work. This is particularly important to the application of automatic speaker recognition systems in the legal setting.

2.2. Automatic speaker recognition systems in forensic speech analysis casework

2.2.1. Forensic Voice Comparison

The vast majority of forensic speech analysis casework consists of Forensic Voice Comparison (FVC). This is the task of comparing two or more recordings to answer questions around

the identities of the speakers within these recordings. Often, the purpose of carrying out FVC is for it to be used as evidence in criminal proceedings. In the UK, evidential FVC is generally carried out by forensic speech science practitioners employing the Auditory-Phonetic and Acoustic (AuPhA) method [13, 14, 15].

It is possible to apply automatic speaker recognition systems to FVC and a substantial amount of research carried out within forensic speech science concerns the use of automatic systems for this task. Indeed, where appropriate, automatic speaker recognition systems are employed (in combination with an AuPhA analysis) in evidential FVC casework in Germany and the Netherlands, for example. However, in UK jurisdictions to date, FVC results derived from automatic speaker recognition systems have not been used as evidence in criminal court proceedings. There are a number of reasons for this, but one which carries particular weight in this respect is the Court of Appeal ruling in R v Slade [2015] EWCA Crim 71.

In R v Slade, the court was presented with voice comparison analysis by way of an automatic speaker recognition system. The appellants sought to rely on this as evidence to support their appeal for conviction. The court voiced a number of concerns about this evidence, and ultimately declined to admit it in that case. One of the concerns related to how evidence of this nature should or could appropriately be presented to a jury. In particular, the court emphasised the need for a jury to understand what aspects of the voices, including the similarity or dissimilarity between them, have contributed to the result produced by the automatic system. The court then, in its reasoning to exclude the evidence, highlighted the importance of explainability. It is worth emphasising that the court did not make a definitive ruling in this case as to whether automatic speaker recognition evidence can ever be admissible, but what is clear is that for this type of evidence to be admitted in future cases, experts need to be able to adequately address the concerns raised by the Court of Appeal, including the issue of explainability.

2.2.2. The issue of explainability

It is understandable that explainability has a special status in the legal setting given that decision makers have to be able to arrive at their conclusions in an informed way. Recognising the importance of effectively communicating forensic science evidence in UK courts, the Royal Society started a project to develop 'judicial primers' which are short documents that aim to explain various forensic analysis methods to judges. Explainability also emerged as an important topic to key legal stakeholders in the US in the work of [16] who carried out semi-structured interviews with forensic laboratory managers, prosecution and defence lawyers, judges and legal academics. The interviews partly focused on how these stakeholders perceived the use of artificial intelligence and machine learning in pattern evidence disciplines (which would include FVC). Some of the participating stakeholders raised concerns around explainability and the prospect of these systems being understood by lay factfinders.

The interview methodology applied by [16] prompted a response from [17] who claim the interview question about the use of artificial intelligence systems was a leading one in that it made assumptions about artificial intelligence not being understandable. [17] argue that the leading question is likely to have introduced bias in the stakeholder responses and, as a result, that this exacerbates the "opacity myth". [17] go on to discuss that while it would be ideal if the trier-of-fact could understand how artificial intelligence systems work, it is very unlikely that they will. They suggest that the trier-of-fact can instead find trust in these systems by understanding that the system being used has been validated using case-relevant data, and that those validation results demonstrate a sufficient level of performance. In effect, [17] argue that there is no need for explainability. The stakeholder responses in [16] appreciated the importance of validating these artificial intelligence systems for use. However, in addition to validation, the stakeholders also recognised the importance of explainability.

Given the type of stakeholders involved in the study in [16] (forensic laboratory managers, lawyers, judges, legal scholars), we are not convinced that a non-leading question would have elicited materially different responses. Furthermore, admissibility of evidence is, in the end, determined by the rules of evidence applicable in any particular jurisdiction. In England and Wales, the applicable law has placed emphasis on the need for explainability (as well as validation) and therefore we, as a community, should continue our efforts to break down the current barrier around explainability.

2.2.3. Addressing explainability

An effort to explain how automatic speaker recognition systems work can be seen in [18], who attempted to identify the types of linguistic units an automatic speaker recognition system exploits most. [18] passed training and test samples, which only consisted of individual phone units, through an i-vector-based automatic speaker recognition system. This was to observe relative performance patterns between phone units. The idea is that if certain phone units yield higher overall system performance compared to others, we may be able to infer what type of linguistic information the system latches on to. Another example of research which could feed into explaining the workings of automatic speaker recognition systems can be seen in [19]. [19] investigated the relationship and degree of complementarity between MFCCs (typically extracted for the purpose of automatic speaker recognition) and voice features a human expert might evaluate as part of FVC. Here, it was long-term formant values and voice quality ratings. Results from [19] may be used in order to infer the type of phonetic information automatic systems capture.

Both [18] and [19] are examples of research which could help to establish a mapping between the input / output of automatic speaker recognition systems and individual linguistic features. As the latter are more readily interpretable by humans, this type of research could address explainability of automatic speaker recognition systems. While research of this type is worthwhile, we need to be alert to the possibility that this, on its own, may not provide us with a sufficiently satisfactory answer. In our view, we need to introduce a different approach that complements the existing research efforts. In addition to explaining automatic speaker recognition results with reference to a componential feature framework, we propose that an approach to explainability which is more holistic should be explored which involves the evaluation of synthetic speech produced by a zero-shot adaptation system. In this new approach, listeners would be played a genuine sample of speech produced by a particular speaker. They would then be played a synthesised sample of that speaker's voice. It would be explained to listeners that the level of similarity between the genuine and the synthesised voices is a result of the synthesis system picking out speaker-specific information from the genuine voice sample. It does so by extracting a holistic model

of that speaker's voice. It is these holistic models of speakers' voices which automatic speaker recognition systems exploit. By allowing listeners to evaluate synthetic speech samples, they can experience what a speaker recognition system does. This 'immersive' experience is then a tool to explainability.

3. Evaluation

Our proposal of using synthetic speech produced by a zero-shot adaptation system in order to explain the workings of an automatic speaker recognition system stems from our experience of working with these speech samples. We are confident, based on our own evaluations and those of others, that zero-shot adaptation can produce synthetic speech which sounds very much like the target speaker. However, in order to increase the credibility of our proposal, a more comprehensive evaluation of this type of synthetic speech may be carried out. Specifically, further perceptual evaluations could take place involving comparisons of the genuine samples of speakers with synthesised samples of those same speakers to see if lay listeners consider them to be the same speaker. Given that a jury panel would be one of the main end-users of our proposal, the evaluations should involve lay listeners who are drawn from a diverse demographic background.

As a further line of evaluation, synthesised and genuine human speech samples could be compared with respect to their acoustic characteristics. Figures 2 – 4 exemplify this line of enquiry using the ASVspoof 2019 dataset. From that dataset, we have taken genuine human speech samples from 48 speakers and compared these to the synthesised speech samples of those 48 speakers in respect of the average long-term formant values for 1st to 4th formants (F1 – F4). As can be seen from these figures, the long-term formant distributions are extremely closely aligned between the genuine human speech samples and the synthesised samples produced by [4]'s system (here, referred to as A10). Instead of focusing on long-term acoustic features, analyses could also target individual vowel and consonant sounds as exemplified in the work of [20]. One purpose of these types of acoustic comparisons would be to see how much speaker-specific acoustic information is retained in the synthesised samples. This, in turn, might also draw out which features of a speaker's voice are captured by an automatic speaker recognition system, and which ones are not.

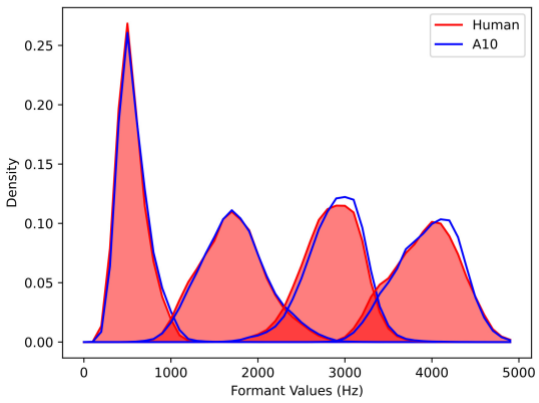

Figure 2. *Probability density plot for long-term formants (F1-F4) for genuine and synthesised speech from all 48 speakers*

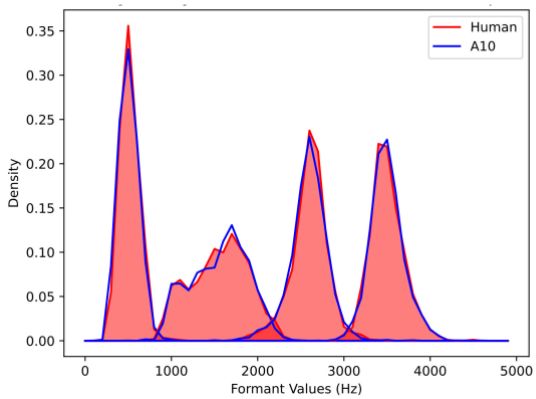

Figure 3. *Probability density plot for long-term formants (F1-F4) for genuine and synthesised speech from a male speaker of Scottish English*

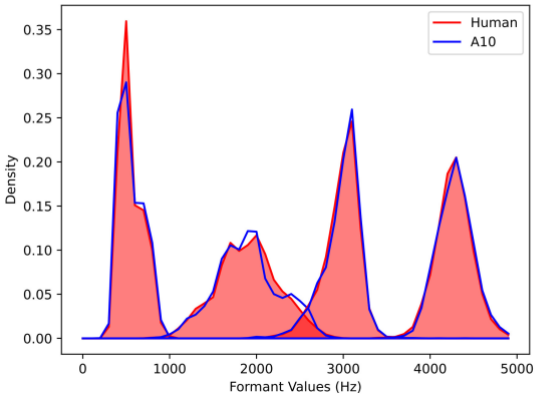

Figure 4. *Probability density plot for long-term formants (F1-F4) for genuine and synthesised speech from a female speaker of Northern British English*

4. Discussion

The primary purpose of the current paper is to propose a new application of evaluating synthetic speech: increasing the explainability of automatic speaker recognition systems. This is particularly valuable for applying them in forensic speech analysis casework. Unlike other types of pattern evidence, speech is in an unusual position whereby it is possible to tap into the senses of the audience. Allowing listeners to experience the degree of similarity between genuine and synthesised speech samples could function as a tool to explaining automatic speaker recognition.

In order to see whether our proposal to explain automatic speaker recognition systems has the potential to meet the expectations of a legal audience, we need to be ready to engage with relevant legal stakeholders. This could take the form of developing a prototype explainability package around automatic speaker recognition systems. A key feature of this explainability package would be the immersive experience made possible by zero-shot adaptation synthetic speech. The explainability package could then be presented to legal professionals in order to elicit their feedback.

5. References

[1] X. Tan, T. Qin, F. Soong, and T. Liu, "A survey on neural speech synthesis," *arXiv preprint*, arXiv:2106.15561. 2021.

[2] A. Black and K. Tokuda, "The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common databases," in *Proc. INTERSPEECH 2005 – Eurospeech, 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, pp. 77-80. 2005.

[3] W.C. Huang, E. Cooper, Y. Tsao, H-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, pp. 4536-4540. 2022.

[4] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," In *Proc 32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, Montreal, Canada, 2018.

[5] D. Snyder, D Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," In *Proc ICASSP 2018 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, pp. 5329-5333. 2018.

[6] E. Cooper, C-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi. Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings. In *Proc ICASSP 2020 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 6184-6188. 2020.

[7] C. Veaux, J. Yamagishi, and K. MacDonald… "CSTR Voice Cloning Toolkit Corpus: English multi-speaker corpus for CSTR voice cloning toolkit. 2017.

[8] V. Panayotov, G.Chen, D.Povey, and S. Khudanpur, "LibriSpeech: An ASR corpus based on public domain audio books. In *Proc ICASSP 2015 – IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*), pp. 5206-5210. 2015.

[9] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," In *Proc. INTERSPEECH 2019 – 20th Annual Conference of the International Speech Communication Association*, Graz, Austria, pp. 1008-1012. 2019.

[10] H. Delgado, N. Evans, T. Kinnunen, K. A. Lee, X. Liu, A Nautsch, J. Patino, M. Sahidullah, M Todisco, X. Wang, and J, Yamagishi, "ASVspoof 2021: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," arXiv:2109.00535. 2021.

[11] C. Terblanche, P. Harrison, and A. Gully, "Human spoofing detection performance on degraded speech," in *Proc. INTERSPEECH 2021 – 22nd Annual Conference of the International Speech Communication Association*, Brno, Czechia, 2021. pp 1738-1742.

[12] C. Kirchhübel and G. Brown, "Spoofed speech from the perspective of a forensic phonetician," in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, 2022, pp. 1308-1312.

[13] M. Jessen, "Forensic voice comparison," in J. Visconti (Ed) *Handbook of Communication in the Legal Sphere*, De Gruyter Mouton, Berlin, pp. 219-255. 2018.

[14] P. Foulkes and P. French, "Forensic speaker comparison: A linguistic-acoustic perspective," In L. M. Solan and P. M. Tiersma (Eds) *Oxford Handbook of Language and Law*, Oxford University Press, Oxford, pp. 557-572. 2012.

[15] C. Kirchhübel, G. Brown and P. Foulkes, "What does method validation look like for forensic voice comparison by a human expert?" *Science & Justice*, vol 63, pp. 251-257. 2023.

[16] H. Swofford and C Champod, "Probabilistic reporting and algorithms in forensic science: Stakeholder perspectives within the American criminal justice system," *Forensic Science International: Synergy*, vol 4, doi: 10.1016/j.fsisyn.2022.100220. 2022.

[17] G. S. Morrison, N. Basu, E. Enzinger, and P. Weber, "The opacity myth: A response to Swofford & Champod (2022)," *Forensic Science International: Synergy*, vol 5, doi: 10.1016/j.fsisyn.2022.100275. 2022.

[18] J. Franco-Pedroso and J. Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol 76, pp. 61-81. 2016.

[19] V. Hughes, P. Harrison, P. Foulkes, P. French, C. Kavanagh, and E. San Segundo, "Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing", in *Proc. INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden, pp. 3892-3896. 2017.

[20] A. Pandey, S. Le Maguer, J. Carson-Berndsen, and N. Harte, "Production characteristics of obstruents in WaveNet and older TTS systems" in *Proc. INTERSPEECH 2022 – 23rd Annual Conference of the International Speech Communication Association*, Incheon, Korea, pp. 2373-2377. 2022