Mind your p's and k's – Comparing obstruents across TTS voices of the Blizzard Challenge 2013

Ayushi Pandey[1], *Sebastien Le Maguer*[1], *Julie Carson-Berndsen*[2], *Naomi Harte*[1]

[1]Sigmedia Lab, ADAPT Centre, School of Engineering, Trinity College Dublin, Ireland
[2]ADAPT Centre, School of Computer Science, University College Dublin, Ireland

pandeya@tcd.ie, lemagues@tcd.ie, julie.berndsen@ucd.ie, nharte@tcd.ie

Abstract

Obstruent consonants have been investigated in speech quality assessment studies of natural speech, where enhancing their perception has improved overall speech quality. This paper presents a comparative analysis of acoustic-phonetic features of obstruent consonants in synthetic speech. Features for obstruent consonants are identified where TTS systems differ significantly from a natural human voice, as a function of quality.

The synthetic speech voices from the Blizzard Challenge of 2013 are used for this investigation. TTS systems were first assigned groups based on their MOS rating (quality) and shared TTS technique (family). Then, acoustic-phonetic features characteristic of contrastive properties in obstruents, were extracted from all systems. While quality differences between low-rated systems and high-rated systems were observed in a large number of features, we report those where statistically significant differences (p-val < 0.001) were observed between the systems. Where quality effects were not found, we investigated whether systems of the same family exhibit similar behaviour. Finally, individual systems within a group were examined for their differing influence on the acoustic-phonetic feature set of obstruents. Here, we found that HMM systems with similar MOS ratings do not differ in their acoustic realization of obstruents, while Unit Selection systems showed stronger individual system variability.

A comparative analysis of obstruent consonants across TTS systems applies techniques from the domain of corpus-phonetics to the task of speech synthesis evaluation. Identifying phonologically relevant acoustic features, may indicate the underlying articulatory process compromised in those systems, that correlates with the distorted acoustics.

1. Introduction

Methods in speech synthesis evaluation have looked at a variety of tools and techniques to analyze synthetic speech in recent years. Some techniques assess the efficacy and scalability in real-world scenarios, like interactive avatar-based settings [1] and long-form, paragraph-level sentences [2]. Objective measurement-based techniques use comparative features like mel cepstral distortion, and the PESQ family of ITU-T standards to predict speech quality compared to a natural voice as reference. Such tests reduce the dependence on expensive human-based listening tests. Machine-learning based techniques such as AutoMOS [3] go even further in modelling human responses and reduce the dependence on parallel natural speech as reference [4, 5]. Electroencephalography (EEG) [6, 7] and pupillometry [8] based measurements have explored the relationship between behavioural/neuronal responses of human participants and synthetic speech quality. To compare the perceived qualities of different TTS techniques, comparative MOS and MUSHRA based perceptual judgements have been conducted [9, 10].

Each of these techniques has advantages - ranging from practical environments, to cost-effective techniques, to contributing to knowledge of quality in speech perception. However, a feature-based comparison of systems using acoustic-phonetic attributes of the signal is largely missing from the discussion.

A central question in the domain of acoustic-phonetics is to identify those features in the signal that can contribute to the perception of contrast between speech sounds. For example, the low-frequency energy region before the release of the consonant allows us to perceive the difference between the utterances "take a pull" and "take a bull". While contrast may not necessarily be the target percept in studies of speech naturalness or quality, contrastive *features* encode rich information about the characteristics of speech sounds. Comparing TTS systems using these features can provide us with insights into system weaknesses, such as poor reproduction characteristics for specific types of consonants.

This paper is the first work we know of that applies techniques from the domain of corpus-phonetics to the task of speech synthesis evaluation. The dataset used for this analysis is the Blizzard Challenge 2013 (BC-2013), which is a single-speaker, parallel database, covering a variety of TTS techniques. Systems of BC-2013, have been grouped on the basis of their shared TTS technique (family) and MOS (quality). Comparative analysis between these groups has been conducted across each obstruent feature, with the original human voice as the reference. The method used is fully automatic, inexpensive and easily reproducible, even at a large scale. We envisage that such an approach can give speech synthesis researchers much greater insights into how the synthetic speech their system produces may be perceived, before conducting subjective evaluation. Features identified in this analysis can be used for comparison between different TTS techniques, system qualities and individual differences between systems.

The paper is organized as follows: Section 2 discusses the properties of obstruent consonants, and the motivation for their choice in this study. Section 3 gives a detailed description of the experimental procedure, entailing the dataset, the feature extraction, and the statistical model. Section 4 presents the results and Section 5 the discussion. Section 6 concludes the paper.

2. Why study obstruents?

Obstruent consonants are a major phonological class of consonants, accounting for 6 distinct phoneme types for stops, [p, t, k, b, d, g], 9 for fricatives, [f, v, θ, ð, s, z, ʃ, ʒ, h], and 2 for affricates [ʧ, ʤ] in English. Obstruents cover a large portion of the consonantal region in any language or dataset. Cross-

	Bilabial		Labiodental		Dental		Alveolar		Postalveolar		Velar		Glottal
Stop	p 122	b 130					t 519	d 402			k 191	g 78	
Affricate									ʧ 35	ʤ 32			
Fricative			f 130	v 122	θ 53	ð 219	s 314	z 172	ʃ 96	ʒ 1			h 218

Table 1: *Frequency distribution of obstruent consonants in the 100 sentences of BC-2013 corpus. Each system has an identical distribution. The rows represent the manners of articulation, while the columns represent the places of articulation.*

linguistic evidence [11] suggests that obstruents cover between two-thirds and three-quarters of the frequency in phoneme inventories across different language groups. In the BC-2013 dataset, obstruents cover 63.9% of the total consonantal population. Their statistical dominance in the dataset makes a compelling case for their analysis.

In addition to their widespread coverage, obstruent consonants have also been evaluated for their contribution to improved speech quality, and poor recognition in noise. In a sequence of studies, Li and Loizou [12–14] report that improved access to obstruents improves intelligibility of speech in noise. Additionally, obstruent recognition has also been found to be more impaired in degraded listening conditions [15, 16], compared to sonorants and vowels, whereas the manipulation of their target cues [17] results in improved recognition. Each of these studies underscore the critical role that preserving obstruents can play in speech perception in non-ideal listening conditions. In this paper, we postulate that synthetic speech may be considered as another such non-ideal scenario. Finally, obstruents contain many acoustic properties of the speech signal, which are not found in sonorants. For instance, stops are characterized by complete obstruction of airflow, which results in a region of silence, followed by a short, high-energy transient region known as burst. Analyzing stops gives us insights into how rapid changes of energy within the acoustic spectrum are handled across different systems. Fricatives do not obstruct the air completely, but force the air through a narrow constriction. This results in air flowing out at high volume velocity, resulting in aperiodic signal with amplitude in high frequencies.

Synthetic speech in BC-2013 contains a range of speech-qualities, and a large proportion of obstruents. Thus we can compare systems in terms of their influence on obstruent properties and explore whether we can uncover relationships with quality that have been established in natural speech. The next section describes the details of BC-2013, our feature extraction procedure and explains the statistical model used for this analysis.

3. Experimental setup

3.1. Dataset

The Blizzard Challenge (BC) is an international task designed to compare state of the art corpus-based speech synthesis systems[1]. All participating teams are given the same training dataset. To participate in the challenge, all teams submit the same prescribed sentences as outputted by a TTS system of their own design. A subset of these sentences are then evaluated with subjective listener tests using MOS.

[1] https://www.synsig.org/index.php/Blizzard_Challenge

In this study, we use data from Blizzard Challenge 2013 (BC-2013). To generate the test sentences, 5 teams used parametric HMM-based techniques (systems C, F, H, I, P), 3 used Unit-Selection (systems B, L, N), and 2 used Hybrid method (systems K,M) for synthesis. Each team submitted the same 100 test sentences, which made BC-2013 a rich source for parallel synthetic speech, with controlled variability.

For the subjective listener test, 11 sentences were evaluated by 426 listeners. While many attributes of speech quality were evaluated, in this work, we focus on the perceived naturalness of the systems. Overall, system M was rated as the most natural and most similar to human speech, with a median MOS of 4 on a 5-point scale. Systems K (Hybrid), I, C (HMM) and L, N (Unit Selection) were the next most highly ranked. System P (HMM) was considered the least natural, and received a MOS of 1.2. In our analysis, the full 100 sentences submitted by each system were used for comparative analysis.

3.2. Feature extraction

This section discusses the feature extraction procedure. First, we discuss the phoneme and sub-phonemic boundary identification in the time domain. Then, we detail the signal processing specifications required for extraction of features from the noisy region of obstruents.

*3.2.1. **Temporal boundary identification***

For phoneme boundary estimation, all systems were forced-aligned using the Montreal Forced Aligner (MFA) [18]. Regions marked for obstruents could now be extracted from the resultant phoneme boundaries. The most important acoustic correlates of obstruent consonants are features extracted from the noisy region of the consonants. While noise continues in fricatives through the length of the consonant, in affricates and stops, it follows a region of silence. Therefore, a sub-phonemic demarcation of the noise region, separated from the silent region needed to be identified.

While most studies on obstruent contrasts depend on careful, hand-corrected methods for the analysis, it would have rendered our corpus-based approaches quite unscalable. Similarly, toolkits such as AutoVOT [19] require a sample of hand-annotated training data, and did not provide the best results for pre-vocalic and intervocalic consonants. However, visually examining the spectrographic properties of stops and affricates, we found a sharp increase in amplitude, representing the burst. To extract this location automatically, we first converted the consonantal signal to its frequency domain. Then, all amplitude values $<$1.5 kHz were removed, because energy from the low-frequency voicing-bar interfered with the estimation of the energy of the burst. Finally, the remaining frequency-domain signal was passed through a moving-average filter. Where en-

ergy of the signal exceeded a threshold of 50-55 dB, and the point of the highest amplitude in that interval was marked as the beginning of the noise region. The threshold was decided upon after examining 20% of the sentences manually.

3.2.2. ***Feature-set***

Acoustic-phonetic properties of obstruents across durational [20–22] , amplitudinal, spectral [23–25] and transitional cues [26–28] are well-established in the literature. The feature extraction procedure closely follows the methodologies presented in Jongman et al.'s seminal work on fricatives [24], and their recent, and more comprehensive extension into all manners of obstruents [29]. The present discussion omits transitional cues and limits the analyses only to the consonantal portion of obstruents. The RMS amplitude has also been calculated in the frequency domain. Also, those cues which cannot be compared across all manners of articulation (for example, closure duration is only relevant for stops and affricates) are excluded.

To extract the spectral parameters, all instances of obstruents were first passed through a high-pass filter, so that the analysis spectrum remains between 550 Hz and 10,000 Hz, to separate source and filter characteristics [30, 31]. For fricatives, a full Hamming window was placed at the center of the frication noise. For stops and affricates, a half Hamming window was placed at the start of the burst, such that the silence region was not included. Then, spectral properties were computed using an 512-point FFT taken over these windowed signals. A brief description is provided below:-

- *Consonant duration:-* The duration of the consonantal region, as returned by the MFA. In the pre-vocalic position, this region starts with the beginning of the closure, and ends with the onset of the vowel. Conversely in the post-vocalic position, it begins at the offset of the vowel, and follows to the end of the consonant. The unit of measurement was milliseconds (ms).
- *Noise duration:-* For stops and affricates, as described above. For fricatives, since noise persists through the length of consonant, the entire region was included. The unit of measurement was milliseconds (ms).
- *RMS amplitude:-* The root-mean-squared amplitude of the power spectrum.
- *Peak amplitude:-* The value of the highest amplitude in the spectrum. The unit of measurement is dB.
- *Peak frequency:-* This is the spectral frequency at which peak amplitude was identified. Its value was measured in Hz.
- *Dynamic amplitude:-* The difference between the peak amplitude, and the minimum amplitude below 2 kHz. The unit of measurement was dB.
- *Spectral tilt:-* The frequency domain of the spectrum was log-transformed, and then a least-squares regression line was fitted through it. The slope of this line returned the spectral tilt.

These features were extracted for obstruent consonants across all the systems, as well as the natural voice, independently. The purpose of such an extraction was to compare these features across all the systems, and to identify those features, where the system (or groups of systems, See Section 3.3) showed significant differences from the natural voice.

R	Group	Sys.	Description
R1	Hybrid-R1	M K	Hybrid systems with MOS 3-4
R2	HMM-R2	I C	HMM systems with MOS 2-3
	UnS-R2	L N	UnS systems with MOS 2-3
R3	HMM-R3	H F	HMM systems with MOS 1-2
	UnS-R3	B	UnS systems with MOS 1-2
R4	HMM-R4	P	HMM systems with MOS 1

Table 2: *Grouping strategy. Rank(R) of the system is decided by MOS for naturalness. The groups correspond to the intersection of the rank and the system family (Hybrid, HMM, Unit Selection (UnS).*

3.3. Grouping strategy

As mentioned in the previous section, the BC-2013 provides a variety of synthetic speech systems, which differ both in family and quality. To achieve this comparative analysis, a grouping strategy between systems was created. The explanation for each of the schemes is described below, and a concise description is displayed in Table 2. Systems were first divided into 4 groups: R1, R2, R3 and R4. R denotes "rank", which was decided simply by the obtained naturalness MOS for a given system. Systems that received MOS in the same interval, i.e, shared the system quality attribute, were assigned the same rank. A comparison based only on rank would not have yielded any family specific insights. Therefore, these groups were further subdivided, so that all systems of the same rank and same family were grouped together. Therefore, the resultant groups were: Hybrid-R1, HMM-R2, UnS-R2, HMM-R3, UnS-R3 and HMM-R4, where UnS means Unit Selection. This strategy allowed us to compare high-rated systems with low-rated systems from the same family. HMM-R4 received poor ratings, and has not been discussed in this paper.

3.4. Statistical model

A linear regression analysis models the relationship between two variables. A linear regression analysis with feature value as the dependent variable, and system group as the predictor variable was conducted for each of the features described in Section 3.2. Separate models were created for each feature, such that the dependent variable changed with every feature in the model, while the independent variable remained system groups each time.

It must be carefully noted here, that the feature value calculated for the natural voice was considered the reference point (the intercept) in each case. The **deviation** from this voice was the comparative metric across which different behaviours of groups were recorded. A univariate analysis of this type allowed for a descriptive model of system group against features, where effect of system groups on each feature could be independently analyzed, and comparative results could be reported.

4. Results

4.1. Experiment I : Comparing the same families of different ranks

The purpose of this experiment is to explore quality differences between groups of the same family. The groups under comparison are HMM-R2 vs HMM-R3, and UnS-R2 vs UnS-R3. Features which showed the most statistically significant differences

between groups have been identified. Comparative influences of groups on such features is presented in the subsequent sections.

4.1.1. ***Comparison between HMM-R2 and HMM-R3***

The most informative features for observing quality differences between HMM-R2 and HMM-R3 were RMS amplitude, peak amplitude and spectral tilt.

On the basis of RMS Amplitude, we see differences between HMM-R2 and HMM-R3 across each manner of articulation. In affricates and fricatives, the HMM-R3 systems were observed to lower the RMS Amplitude. HMM-R2, on the other hand, did not differ significantly from the natural voice in any manner of articulation. RMS Amplitude dropped in affricates by 1.8 dB, and in fricatives by 1.5 dB, with strongly significant effects (p-val < 0.001). In stops, HMM-R3 systems were found to increase the amplitude by 0.51 dB, with a moderately significant effect (p-val < 0.05). Therefore, through these results we can conclude that poor-quality HMM-R3 systems show lower amplitude in affricates and fricatives, and marginally higher amplitude compared to natural voice. In each case, HMM-R2 was not found significantly different from natural voice.

The second feature under consideration is the peak amplitude. Similarly as above, HMM-R3 systems are found to lower the peak amplitude in the context of affricates and in fricatives. The peak amplitude dropped in affricates by 2.4 dB, and in fricatives by 1.4 dB, with significant effects (p-val < 0.01). HMM-R2 systems, on the other hand, do not differ from the natural voice in affricates. On the contrary, they are seen to increase the amplitude for fricatives. The behaviour of the two groups was not different in stops. Therefore, we can learn that fricatives in HMM-R2 systems exhibit louder maxima of amplitude, and HMM-R3 have softer peak amplitudes in affricates and fricatives alike.

The third feature considered important is the spectral tilt. In all the manners of articulation, low-quality HMM-R3 systems increase the spectral tilt with strongly significant effects. The magnitude of this increase is 1.93 dB in affricates, 4.14 dB in fricatives, and 3.14 dB in stops (p-val < 0.001). In affricates and fricatives, HMM-R2 systems do not differ significantly from the natural voice. But in stops, HMM-R2 also increase the spectral tilt. However, groups can still be separable within this context, because the magnitude of this increase is much lesser (0.95 dB) than in HMM-R3. Therefore, we observe that fricatives and affricates have steeper slopes in low-quality HMM systems across all manners of articulation. But in the context of stops, HMM-R2 also contribute to this effect.

4.1.2. ***Comparison between UnS-R2 and UnS-R3***

The most important features for comparison between UnS groups are consonant duration, noise duration and spectral tilt.

Both UnS-R2 and UnS-R3 systems shorten the consonant duration in the context of fricatives and stops, while affricates do not show differences in groups for consonant duration. However, the shortening in high-quality UnS-R2 systems is seen with a stronger effect (p-val < 0.001), compared to UnS-R3 systems. In UnS-R2, fricatives are shortened by 7.5 ms and stops by 5.8 ms. In UnS-R3, on the other hand, fricatives and stops are shortened by 4.4 ms and 2.6 ms, respectively (p-val < 0.01). Therefore, we observe here that high-quality UnS-R2 systems shorten fricatives and stops more than low-quality UnS-R3.

The second feature considered important for UnS quality comparison is noise duration. Similar to observations for noise duration, a decrease of noise duration is found in both UnS-R2 and UnS-R3 groups for all manners of articulation. However, there are two differences. Firstly, stops show comparable decrease of noise duration between UnS-R2 and UnS-R3, and therefore are not deemed a reliable context for group differentiation. Secondly, although both fricatives and affricates have different influences of groups, they do so in different directions. UnS-R2 systems reduce the duration of fricatives with stronger significance, but affricates are shortened in UnS-R3 more strongly. Fricatives in UnS-R2 are shortened by 7.5 ms (p-val < 0.001), compared to 4.4 ms in UnS-R3 (p-val < 0.01). On the other hand, affricates are shorter by 7.4 ms in UnS-R2 (p-val < 0.05), and 9.8 ms (p-val < 0.01) in UnS-R3. So here, we can learn that noise duration is reduced in both UnS-R2 and UnS-R3 groups, across all manners of articulation. Group differences can be seen within fricatives and affricates. But the direction of influence is not consistent across manners.

The third feature under consideration is the spectral tilt. Here we see, that UnS systems on the whole lower the spectral tilt, instead of the increasing effect found in HMM systems. While the effect of lowering is strong and significant in all manners of articulation alike (p-val < 0.001), affricates and fricatives show greater separation between UnS-R2 and UnS-R3. In affricates, UnS-R2 decrease the tilt by 3.3 dB, and UnS-R3 by 7.3 dB. Similarly for fricatives, UnS-R2 decrease the tilt by 5.43 dB, and UnS-R3 by 8.7 dB. Stops, on the other hand, show comparable lowering in both UnS-R2 and UnS-R3 groups. Therefore, this result indicates that low-quality UnS-R3 systems flatten the spectral tilt more than UnS-R2 system, especially for fricatives and affricates.

4.2. Experiment II : Comparing individual differences between systems of a group

The purpose of this experiment is to explore individual differences between systems of the same group. Comparison will be made under Hybrid-R1 between M and K, under HMM-R2 between I and C, and under UnS-R2 between L and N.

4.2.1. ***Comparison between individual systems of Hybrid-R1***

It is important to note that although M and K are in the same group, with obtained MOS of 3.9 and 3.4 respectively, that difference was statistically significant in the BC-2013 evaluations. The three most important features identified for systemic differences are RMS amplitude, peak frequency and spectral tilt.

Regarding **RMS Amplitude**, in the context of affricates, M was found to lower the RMS Amplitude by 1.7 dB (p-val $<$ 0.001), but K was not found to be significantly different from the natural voice. However, this trend completely reversed in the context of fricatives and stops. K was observed to influence a strongly significant increase the amplitude of 1.72 dB (p-val < 0.001). But in both of these contexts, M was not found different from the natural voice. Therefore, affricates are softer than natural voice in M, and fricatives and stops are louder in K. So we can see that, although each manner of articulation shows systemic differences between Hybrid systems, affricates oppose the trend exhibited by fricatives and stops.

The second feature considered reliable for systemic differences within Hybrid-R1 is **peak frequency**. K shows a statistically significant raising of peak frequency in all affricates, fricatives and stops context. In affricates, the increase is by 946.23 Hz, while in fricatives, we see an increase of 337.46 Hz. Finally in stops, although the increase is smallest, of 201.8 Hz compared to other places, the effect is still strongly significant. In no context does M differ from the natural voice. Therefore,

K exhibits maximum amplitude at higher frequencies, while M remains closer to natural.

Finally, K shows a statistically significant raising of spectral tilt in each context. The increase was of 1.2 dB in affricates, 5.4 dB in fricatives, and 3.5 dB in stops. M does not differ significantly from the natural voice in fricatives and stops. However, greater separation in systems can be seen in affricates, where M shows a moderately significant lowering of the spectral tilt (p-val < 0.05). Therefore, K shows a steeper slope in the spectrum, while M does not differ significantly from the natural voice.

*4.2.2. **Comparison between individual systems of HMM-R2***

Differences between I and C were **not found** in any feature, across any manner of articulation. This indicates that systems I and C have consistent patterns of influence on all the features across manners of articulation.

*4.2.3. **Comparison between individual systems of UnS-R2***

The first feature to compare differences between L and N is **RMS Amplitude**. Differences on the basis of RMS Amplitude can be seen in all three classes of Manner - i.e., in affricates, fricatives and stops. In affricates and fricatives, N shows a strongly significant lowering of RMS Amplitude. The magnitude of this lowering is 3.0 dB and 2.9 dB in affricates and fricatives respectively (p-val < 0.001). L, on the other hand, does not differ significantly from the natural voice. Among stops, the difference is less distinct, because N brings about only a modest lowering of 0.56 dB (p-val < 0.05).

The second feature under consideration is **peak frequency**. Systemic differences can be seen predominantly in affricates, and modestly in Stops. In affricates, L shows a moderately significant lowering of 211.86 Hz (p-val < 0.05), while N does not differ much from the natural voice. Among stops, although the systems differ individually, the pattern of affricates is not replicated. Here, both L and N show a lowering of the frequency. The effect although, is stronger in N, with a lowering of 173.14 Hz (p-val < 0.001), compared to L which lowers by 142.76 Hz (p-val < 0.01).

Finally, differences based on **spectral tilt** can be seen in all three classes of Manner. In affricates and stops, N shows a strongly significant lowering of 5.55 dB (p-val < 0.001) and 3.15 db (p-val < 0.001) respectively, and L does not differ from the natural voice. In fricatives, the difference between systems is less clearer, because both N and L show lowering. However, a greater magnitude of lowering can be observed in N, of 8.7 dB with a strongly significant effect.

5. Discussion

In the previous section, we saw a detailed description of results gathered from the two experiments. Spectral tilt can clearly be seen to show important differences for each of the phenomena under consideration. From Experiment I, it can be seen that HMM-R3 show increased spectral tilts, while HMM-R2 do not differ significantly. Similarly, in Experiment II, comparatively lower-rated K, and N showed increased spectral tilts, in the Hybrid-R1 and UnS-R2 groups, respectively. This is consistent with previous findings on flatter spectral tilt contributing to improved intelligibility [32]. Although there is little agreement on the relationship between naturalness and intelligibility, we find that spectral tilt appears to differentiate system-groups based on naturalness as well.

System-family specific results can also be observed on the basis of spectral tilt, and on consonantal duration. In HMM-R3 systems, spectral tilt increases from the natural voice. However, in low-quality UnS-R3 systems, it is seen to decrease more steeply. Therefore, spectral tilt exhibits quality-specific differences, but the influence is family-dependent. In terms of perceived speech quality, this indicates a preference for preserving the spectral tilt, and that deviation in either direction compromises quality.

Another important result can be seen is that UnS systems show differences based on quality in *durational* cues, while HMM systems on the other hand, impact spectral features more. It may be speculated here that statistical averaging practised in HMM systems, compromises the necessary variation required to retain spectral features. From these results, we can also speculate that the cost function of the unit selection systems favors shorter units over longer ones. A deeper investigation about which units have been selected would bring a better insight about the reason of this trend.

Finally, from Experiment II, we see important individual variation between UnS-R2 systems, and none whatsoever between HMM-R2 systems. While systems of HMM-R2 are more closely rated in naturalness and intelligibility, UnS-R2 have also received quite similar ratings [33]. Therefore, good-quality HMM systems rigidly approach statistical averaging and filter out variation between systems.

6. Conclusion

In this study, we have presented a comparative analysis of TTS systems from the BC-2013, using acoustic-phonetic measurements extracted from obstruent consonants. 10 systems from BC-2013 were grouped on the basis of their quality and family. A linear regression analysis was conducted to establish a relationship between system groups and acoustic measurements, with the natural voice as reference. Spectral tilt emerged as the most informative feature, where several different phenomena of quality, family and individual system differences could be observed. In general, better-rated systems were found to be associated with flatter spectral tilts, and higher RMS amplitude values for obstruents. These results were consistent with previous studies on improved intelligibility.

Avoiding the use of expensive behavioural equipment, we have been able to connect the domains of phonetics and speech technology. We have shown that the use of phonetic measurements is useful for a variety of comparison tasks, and the results are meaningful from a speech production and perception standpoint. For future work, we will incorporate transitional cues from adjacent vowels to gain deeper insights into the obstruent behaviour across different systems, especially for analyzing their concatenative ability. The dataset from BC-2013 will be extended to include neural voices built using systems such as Tacotron [34] and FastPitch [35]. A long-term goal of this approach is to identify more acoustic-phonetic features across different phonetic segments, including non-obstruent consonants, vowels and dipthongs.

A complete description of segmental properties of parallel synthetic speech can give speech synthesis researchers immediate feedback about the expectation of naturalness in their systems. These studies can precede subjective evaluation tests, by informing speech technologists about signal distortion at a segment and co-articulation level. Finally, from an acoustic-phonetic point of view, these studies allow us to understand phonemic properties that remain intact in the signal, despite a loss in naturalness.

7. Acknowledgements

This research has the financial support of Science Foundation Ireland under Grant number 18/CRT/6224. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund.

8. References

[1] J. Mendelson and M. P. Aylett, "Beyond the listening test: An interactive approach to TTS evaluation." in *International Conference on Speech Communication and Technology (Interspeech)*, 2017, pp. 249–253.

[2] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," *arXiv preprint arXiv:1909.03965*, 2019.

[3] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. W. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *ArXiv*, vol. abs/1611.09207, 2016.

[4] F. Hinterleitner, *Quality of Synthetic Speech: Perceptual Dimensions, Influencing Factors, and Instrumental Assessment*. Springer, 2017.

[5] S. wei Fu, Y. Tsao, H.-T. Hwang, H.-M. Wang *et al.*, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *International Conference on Speech Communication and Technology (Interspeech)*, 2018.

[6] H. Maki, S. Sakti, H. Tanaka, and S. Nakamura, "Quality prediction of synthesized speech based on tensor structured EEG signals," *PloS one*, vol. 13, no. 6, 2018.

[7] I. H. Parmonangan, H. Tanaka, S. Sakti, S. Takamichi, and S. Nakamura, "Speech quality evaluation of synthesized japanese speech using EEG," *International Conference on Speech Communication and Technology (Interspeech)*, pp. 1228–1232, 2019.

[8] A. Govender and S. King, "Using pupillometry to measure the cognitive load of synthetic speech," in *International Conference on Speech Communication and Technology (Interspeech)*, 2018, pp. 2838–2842.

[9] M. Cohn and G. Zellou, "Perception of concatenative vs. neural text-to-speech (tts): Differences in intelligibility in noise and language attitudes," in *International Conference on Speech Communication and Technology (Interspeech)*, 2020, pp. 1733–1737. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1336

[10] T. Merritt, B. Putrycz, A. Nadolski, T. Ye, D. Korzekwa, W. Dolecki, T. Drugman, V. Klimkov, A. Moinet, A. Breen *et al.*, "Comprehensive evaluation of statistical speech waveform synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 325–331.

[11] B. Lindblom and I. Maddieson, "Phonetic universals in consonant systems," *Language, speech and mind*, vol. 6278, 1988.

[12] N. Li and P. C. Loizou, "The contribution of obstruent consonants and acoustic landmarks to speech recognition in noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 6, pp. 3947–3958, 2008.

[13] ——, "Factors affecting masking release in cochlear-implant vocoded speech," *The Journal of the Acoustical Society of America*, vol. 126, no. 1, pp. 338–346, 2009.

[14] ——, "Masking release and the contribution of obstruent consonants on speech recognition in noise by cochlear implant users," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1262–1271, 2010.

[15] S. A. Phatak, A. Lovitt, and J. B. Allen, "Consonant confusions in white noise," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 1220–1233, 2008.

[16] J. Meyer, L. Dentel, and F. Meunier, "Speech recognition in natural background noise," *PloS one*, vol. 8, no. 11, p. e79279, 2013.

[17] F. Li and J. B. Allen, "Manipulation of consonants in natural speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 496–504, 2011.

[18] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *International Conference on Speech Communication and Technology (Interspeech)*, 2017, pp. 498–502.

[19] M. Sonderegger and J. Keshet, "Automatic measurement of voice onset time using discriminative structured prediction," *The Journal of the Acoustical Society of America*, vol. 132, no. 6, pp. 3965–3979, 2012.

[20] T. Cho and P. Ladefoged, "Variation and universals in vot: evidence from 18 languages," *Journal of phonetics*, vol. 27, no. 2, pp. 207–229, 1999.

[21] B. H. Repp, "Closure duration and release burst amplitude cues to stop consonant manner and place of articulation," *Language and speech*, vol. 27, no. 3, pp. 245–254, 1984.

[22] A. Jongman, "Duration of frication noise required for identification of english fricatives," *The Journal of the Acoustical Society of America*, vol. 85, no. 4, pp. 1718–1725, 1989.

[23] E. Chodroff and C. Wilson, "Burst spectrum as a cue for the stop voicing contrast in american english," *The Journal of the Acoustical Society of America*, vol. 136, no. 5, pp. 2762–2772, 2014.

[24] A. Jongman, R. Wayland, and S. Wong, "Acoustic characteristics of english fricatives," *The Journal of the Acoustical Society of America*, vol. 108, no. 3, pp. 1252–1263, 2000.

[25] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *The Journal of the Acoustical Society of America*, vol. 64, no. 5, pp. 1358–1368, 1978.

[26] H. M. Sussman, H. A. McCaffrey, and S. A. Matthews, "An investigation of locus equations as a source of relational invariance for stop place categorization," *The Journal of the Acoustical Society of America*, vol. 90, no. 3, pp. 1309–1325, 1991.

[27] H. M. Sussman, D. Fruchter, and A. Cable, "Locus equations derived from compensatory articulation," *The Journal of the Acoustical Society of America*, vol. 97, no. 5, pp. 3112–3124, 1995.

[28] D. T. P. D. McCarthy, "The acoustics of place of articulation in english plosives," Ph.D. dissertation, Newcastle University, 2019.

[29] C. Redmon, "Lexical acoustics: Linking phonetic systems to the higher-order units they encode," *PhD dissertation, University of Kansas, Lawrence*, 2020.

[30] C. H. Shadle and S. J. Mair, "Quantifying spectral characteristics of fricatives," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1521–1524.

[31] L. L. Koenig, C. H. Shadle, J. L. Preston, and C. R. Mooshammer, "Toward improved spectral measures of/s: Results from adolescents," 2013.

[32] Y. Lu and M. Cooke, "The contribution of changes in f0 and spectral tilt to increased intelligibility of speech produced in noise," *Speech Communication*, vol. 51, no. 12, pp. 1253–1262, 2009.

[33] S. King and V. Karaiskos, "The blizzard challenge 2013," in *The Blizzard Challenge Workshop*, 2013, http://festvox.org/blizzard/bc2013/summary_Blizzard2013.pdf.

[34] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions," in *international Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[35] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," *arXiv preprint arXiv:2006.06873*, 2020.