Measuring the Perceptual Effects of Modelling Assumptions in Speech Synthesis Using Stimuli Constructed from Repeated Natural Speech

Measuring the perceptual effects of modelling assumptions in speech synthesis using stimuli constructed from repeated natural speech

Gustav Eje Henter[1], *Thomas Merritt*[1], *Matt Shannon*[2], *Catherine Mayo*[1], *Simon King*[1]

[1]Centre for Speech Technology Research, University of Edinburgh, U.K.
[2]Department of Engineering, University of Cambridge, U.K.

ghenter@inf.ed.ac.uk, t.merritt@ed.ac.uk, sms46@eng.cam.ac.uk

Abstract

Acoustic models used for statistical parametric speech synthesis typically incorporate many modelling assumptions. It is an open question to what extent these assumptions limit the naturalness of synthesised speech. To investigate this question, we recorded a speech corpus where each prompt was read aloud multiple times. By combining speech parameter trajectories extracted from different repetitions, we were able to quantify the perceptual effects of certain commonly used modelling assumptions. Subjective listening tests show that taking the source and filter parameters to be conditionally independent, or using diagonal covariance matrices, significantly limits the naturalness that can be achieved. Our experimental results also demonstrate the shortcomings of mean-based parameter generation.

Index terms: speech synthesis, acoustic modelling, stream independence, diagonal covariance matrices, repeated speech

1. Introduction

Statistical parametric speech synthesis (SPSS) has inadequate naturalness (see, e.g, [1]). This is true when sampling from the statistical models [2], which sounds warbly and bubbly—showing the models are poor descriptions of the distribution of natural speech—as well as for maximum likelihood[1] parameter generation (MLPG) [3], which sounds buzzy and muffled.

Whilst neither text analysis nor speech parameter representation are perfect (see, e.g., [4]), the acoustic models are more often cited as the cause because they make various assumptions that, if not true for natural speech, may limit naturalness.

A compelling way to try to improve naturalness is to create models that eliminate problematic assumptions: accurate models should generate better output. But which assumptions adversely affect naturalness? And what are the limits of using ever more accurate acoustic models under different parameter generation methods? Much previous work has proposed new models which mitigate particular assumptions, without first demonstrating that they are actually problematic. For example, *semi-tied covariance matrices* [5, 6] relax conditional independence assumptions between parameters, but we cannot conclude whether those assumptions impose a fundamental limit on naturalness, if other aspects of the model were more accurate.

This paper takes a new and different approach, investigating modelling assumptions by using natural speech of the same text spoken multiple times. These can be seen as conditionally independent samples from a perfect acoustic model of the text. We manipulate and combine natural parameter sequences in ways that, if a particular assumption is correct, will not affect the resulting speech. By measuring how naturalness changes as we perform further manipulations, we gain insight into the relative severity of the corresponding modelling assumptions.

In addition to our previous work in synthesis [7], there has been work in recognition. Some experiments [8, 9, 10] tease apart the relative importance of the language model and acoustic model in human recognition by preventing listeners from using long-range linguistic context, which is not captured well by statistical language models. This tests the fundamental limits of acoustic modelling. Other experiments involving resampling from empirical distributions [11, 12] investigate the severity of conditional independence assumptions made in a typical recognition system. Their manipulations of natural speech parameter sequences are similar to ours, but the methodology is quite different, and the results do not necessarily apply to synthesis.

[1]Although in fact MLPG does not actually maximise the *likelihood*.

2. Background

We begin by outlining relevant aspects of statistical parametric speech synthesis. Detailed introductions are available elsewhere [13, 14]. The main idea of parametric speech synthesis is to use a *vocoder* to represent speech as a sequence $\underline{\boldsymbol{x}} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$ of speech parameter vectors $\boldsymbol{x}_t$, rather than a waveform. Typically, these vectors have high dimensionality—50, or more—to prevent quality loss. Vocoders generally use a *source-filter representation* [15], with each vector $\boldsymbol{x}_t = [\boldsymbol{x}_t^{\mathrm{s}}; \boldsymbol{x}_t^{\mathrm{f}}]$ in the sequence composed from *source* characteristics $\boldsymbol{x}_t^{\mathrm{s}}$ and elements $\boldsymbol{x}_t^{\mathrm{f}}$ representing the *filter* around time t. The filter parameters are commonly *mel cepstral coefficients* (MCEPs).

In SPSS, a conditional probabilistic model $\mathbb{P}(\underline{\boldsymbol{x}} \,|\, l; \nu, \lambda)$ is defined over the speech parameter sequence $\underline{\boldsymbol{x}}$ given the text l [13, 14]. In *training*, the parameters ν and λ of this model are estimated from a corpus of (text, speech parameter sequence) pairs from a human speaker. The trained model can then be used with a *speech parameter generation* method such as sampling or MLPG to generate a parameter sequence for input text.

Typically the model $\mathbb{P}(\underline{\boldsymbol{x}} \,|\, l; \nu, \lambda)$ comprises a *state transition model* $\mathbb{P}(\underline{q} \,|\, l; \nu)$ and an *acoustic model* $\mathbb{P}(\underline{\boldsymbol{x}} \,|\, \underline{q}; \lambda)$, where the *state sequence* $\underline{q} = [q_1, \ldots, q_T]$ is a latent variable. The state q_t consists of phonemic and linguistic context that potentially influences the speech parameter sequence around time t.

Current acoustic models make many assumptions, such as:

1. The source $\underline{\boldsymbol{x}}^{\mathrm{s}}$ and filter $\underline{\boldsymbol{x}}^{\mathrm{f}}$ portions of the speech parameter sequence are conditionally independent given $\underline{q}$:
$$\mathbb{P}(\underline{\boldsymbol{x}} \,|\, \underline{q}; \lambda) = \mathbb{P}(\underline{\boldsymbol{x}}^{\mathrm{s}} \,|\, \underline{q}; \lambda) \cdot \mathbb{P}(\underline{\boldsymbol{x}}^{\mathrm{f}} \,|\, \underline{q}; \lambda) \qquad (1)$$
2. The time trajectories of different mel cepstral coefficients $\underline{x}^{\mathrm{f},m}$ are conditionally independent given $\underline{q}$:
$$\mathbb{P}(\underline{\boldsymbol{x}}^{\mathrm{f}} \,|\, \underline{q}; \lambda) = \prod_m \mathbb{P}(\underline{x}^{\mathrm{f},m} \,|\, \underline{q}; \lambda) \qquad (2)$$

3. The distributions $\mathbb{P}(\underline{\boldsymbol{x}}^{\mathrm{s}} \mid \underline{q}; \lambda)$ and $\mathbb{P}(\underline{\boldsymbol{x}}^{\mathrm{f},m} \mid \underline{q}; \lambda)$ over trajectories given a state sequence $\underline{q}$ are Gaussian.
4. The Gaussian distributions have a particular parametric form, e.g., *trajectory* [16] or *autoregressive* [17] HMMs.

These assumptions are hierarchical: each implicitly assumes the previous ones. The training model usually also makes *frame-wise* conditional independence assumptions; the model used for parameter generation does not (a known inconsistency [14]).

The assumptions in equations (1) and (2) are useful in practice as they simplify the probabilistic model, typically reduce the number of parameters, and tend to make training more computationally tractable. However, they may also limit naturalness. Our goal is to assess the impact of such assumptions.

3. Investigating the limitations of accurate acoustic models by using repeated speech

We use manipulated natural speech parameter sequences to investigate the fundamental limits on naturalness of synthetic speech by using ever more accurate acoustic models. To explore the limitations, we consider a hypothetical, extremely accurate model of speech, called $\mathcal{M}_{\mathrm{D}}$, along with a set of other models, $\mathcal{M}_{\mathrm{SF}}$–$\mathcal{M}_{\mathrm{I}}$, that make certain conditional independence assumptions but otherwise are extremely accurate. The models and their assumptions are detailed in section 5.1. By stitching together natural speech parameter trajectories from multiple examples of the same text being read aloud, we can approximate (up to issues with duration) what speech sampled from each of these different models would sound like.

$\mathcal{M}_{\mathrm{SF}}$ makes the source-filter independence assumption in equation (1), but is otherwise highly accurate. By creating a chimeric[2] speech parameter sequence where the source parameters come from one repetition of a sentence and the filter parameters come from another, we approximate parameter trajectories sampled from this model: The source and filter portions of the parameter sequences are independent in the chimeric speech, but each portion is internally highly consistent and follows the same marginal distribution as natural speech. If the independence assumption in equation (1) is correct, this manipulation will not affect the naturalness of the resulting speech.

Similarly, for sampling from a model $\mathcal{M}_{\mathrm{I}}$ that assumes that time trajectories of all different mel cepstral coefficients are conditionally independent, as in equation (2), but otherwise is extremely accurate, we create a chimeric speech parameter sequence where the trajectory of the 0^{th} mel cepstral coefficient is taken from one repetition, the 1^{st} mel cepstral coefficient from another, etc. Further, by taking the mean of the entire database rather than combining individual examples, we can approximate what mean-based parameter generation (as opposed to sampling) would sound like under models $\mathcal{M}_{\mathrm{D}}$–$\mathcal{M}_{\mathrm{I}}$. Since the mean is unaffected by our independence assumptions, this is the same for all models.

We use a corpus of repeated natural speech (section 4) where the phonemic and linguistic context is the same for every repetition of the text, eliminating acoustic variations due to heterogeneous context. We treat different repetitions of a given prompt as conditionally independent, given the text and the speaker. It is however necessary to accommodate differences in timing between repetitions. SPSS typically accounts for timing variations by introducing a latent state sequence and a duration model (section 2). We use a simple alternative, *dynamic time warping* (DTW), to align different repetitions of a given prompt to a common reference before combining them.[3] The DTW takes source and reference speech parameter sequences, computes an alignment between frames by minimising the *mel cepstral distortion* (MCD) excluding the 0^{th} cepstral coefficient, then generates a time-warped version of the source by dropping or repeating frames as necessary in order to respect the alignment. This imposes the timings of the reference on the source.

[2] A *chimera* is something composed of disparate parts.

4. The REHASP 0.5 corpus

The corpus of repeated speech we created for our investigation uses *prompts* selected from the Harvard sentences [18], which are widely used among phoneticians and speech technology researchers. The Harvard sentences are in sets of ten, each set being approximately phonetically balanced. We selected three sets suitable for British English speakers, yielding 30 prompts.

We recorded a female British English speaker, "Lucy," speaking each of the 30 prompts 40 times in a hemi-anechoic chamber. We term each such recording a *repetition* of a prompt. The prompts were spoken in a neutral style and the speaker was instructed not to intentionally vary the repetitions. To prevent list effects, prompts were presented in random order, except that the same prompt never appeared twice in a row.

This yielded 1200 utterances recorded at 16 bit, 96 kHz. A high-pass filter with 6 dB point at 30 Hz was applied to reject minor low frequency electrical interference. Utterances were then normalised to -24 dBov [19]. The result is the *REpeated HArvard Sentence Prompts* (REHASP) corpus version 0.5. This corpus is publicly and freely available under a permissive license from the Edinburgh DataShare archive at permanent URL http://hdl.handle.net/10283/561.

5. Experiments

To explore the perceptual effects of different assumptions and parameter generation methods, we conducted two listening tests: one measured naturalness; the other examined the dimensions that listeners used to discriminate amongst conditions. Our speech parameter representation and vocoder was "legacy" STRAIGHT [20], with 40 mel cepstral coefficients (MCEPs orders 0 through 39), log fundamental frequency (LF0), and 5 band aperiodicity coefficients (BAPs), at 5 ms frame shift. The REHASP 0.5 corpus was downsampled to 16 kHz, a common operating point for TTS systems.

5.1. Conditions

We investigated the 12 *conditions* in table 1. Conditions N through D are baselines providing context for subsequent comparisons. For condition N the natural waveform for repetition a is used as the generated waveform ($1 \leq a \leq 40$ chosen uniformly at random for each prompt). For condition VU, the natural waveform of repetition a was simply converted to the parameter domain and back (analysis-synthesis). This measures the loss incurred by the parametric representation. Informal listening suggested that certain vocoding artefacts could be removed by applying minor temporal smoothing (Gaussian window; $\sigma = 0.8$ frames) after STRAIGHT-based analysis; this is condition V and is used as the basis for all remaining conditions.

DTW (section 3) was used to align different repetitions of each prompt. For condition D, the speech parameter sequence for repetition a is time warped using repetition b as the reference

[3] The DTW code used is available as a python package at https://pypi.python.org/pypi/mcd/0.2.

Condition				Parameter trajectory sources					
				Dur-	Source		Filter (MCEPs)		
ID	Description	Model	Generation	ation	LF0	BAP	0–5	6–12	13–39
N	Natural speech			-	-	-	-	-	-
VU	Vocoded (unsmoothed parameters)			a	a	a	a	a	a
V	Vocoded (smoothed parameters)			a	a	a	a	a	a
D	Time-warped to reference duration	$\mathcal{M}_{\mathrm{D}}$	Sampling	b	a	a	a	a	a
SF	Source and filter independent	$\mathcal{M}_{\mathrm{SF}}$	Sampling	b	a	a	c	c	c
SI	All parameter streams independent	$\mathcal{M}_{\mathrm{SI}}$	Sampling	b	a	d	c	c	c
L1	Lower 6 MCEPs independent	$\mathcal{M}_{\mathrm{L1}}$	Sampling	b	a	a	$*$	c	c
L2	Lower 13 MCEPs independent	$\mathcal{M}_{\mathrm{L2}}$	Sampling	b	a	a	$*$	$*$	c
H1	MCEPs above 12 independent	$\mathcal{M}_{\mathrm{H1}}$	Sampling	b	a	a	c	c	$*$
H2	MCEPs above 5 independent	$\mathcal{M}_{\mathrm{H2}}$	Sampling	b	a	a	c	$*$	$*$
I	All MCEPs independent	$\mathcal{M}_{\mathrm{I}}$	Sampling	b	a	a	$*$	$*$	$*$
M	MCEPs averaged	$\mathcal{M}_{\mathrm{D}}$–$\mathcal{M}_{\mathrm{I}}$	Mean	b	a	a	$\overline{x}$	$\overline{x}$	$\overline{x}$

Table 1: *The conditions investigated, their corresponding model and generation method, and their construction. An asterisk means that all coefficients were taken from independent, distinct repetitions, while $\overline{x}$ denotes that an average over the entire database was used.*

($1 \leq b \leq 40$ chosen uniformly at random for each prompt; $b \neq a$). This evaluates the degradation due to time warping and provides a baseline for subsequent conditions. We can think of D as sampling from an extremely accurate model $\mathcal{M}_{\mathrm{D}}$.

Conditions SF through I use the methodology described in section 3. Condition SF investigates the source-filter independence of model $\mathcal{M}_{\mathrm{SF}}$, taking source parameter sequence $\underline{\boldsymbol{x}}^{\mathrm{s}}$ from repetition a and filter parameter sequence $\underline{\boldsymbol{x}}^{\mathrm{f}}$ from repetition c ($1 \leq c \leq 40$ chosen uniformly at random for each prompt; $c \notin \{a, b\}$). Condition SI further assumes that source streams LF0 and BAP are independent, representing sampling from model $\mathcal{M}_{\mathrm{SI}}$. These are common assumptions in SPSS.

We also investigated models $\mathcal{M}_{\mathrm{L1}}$ through $\mathcal{M}_{\mathrm{I}}$ making various conditional independence assumptions between the trajectories $\underline{x}^{\mathrm{f},m}$ of different mel cepstral coefficients m. Starting from condition SF, progressively larger blocks of lower-order MCEP trajectories were treated as independent (conditions L1, L2 and I), and similarly for blocks of higher-order MCEP trajectories (conditions H1, H2 and I). For condition I, all MCEP trajectories were treated as independent, which corresponds to using diagonal covariance matrices in Gaussian models.

Finally we considered averaging the aligned filter parameter sequences over all 40 repetitions of each prompt, yielding condition M. This corresponds to mean-based parameter generation with any of the models $\mathcal{M}_{\mathrm{D}}$ through $\mathcal{M}_{\mathrm{I}}$.

5.2. Naturalness evaluation

We first measured the subjective naturalness of the different conditions. Because differences may be subtle, we employed a MUSHRA (*MUltiple Stimuli with Hidden Reference and Anchor*) methodology [21], as used to evaluate audio codecs.

In MUSHRA, subjects rate a set of stimuli in parallel from 0 (very poor) to 100 (completely natural). The stimuli for different conditions are matched (same prompt for all conditions). There is one rating slider for each condition; conditions are ordered randomly and presented without labels. An unprocessed reference stimulus (condition N) is accessible beside the sliders, and also included among the unlabelled examples; subjects are instructed to rate this hidden reference as completely natural, thus fixing the high end of the scale. For audio codecs, the low end is anchored by a 3.5 kHz low-pass filtered version of the reference. While this will have poor signal quality, it is fundamentally natural, being neither vocoded nor modelled. This makes it unsuitable for our test, and it was thus omitted.

30 native speakers of English were asked to rate 20 sets of 12 stimuli, each set presenting the same prompt manipulated according to the 12 different conditions. The 20 prompts selected for each listener represented two out of the three Harvard sentence blocks. The design was balanced such that each of the three blocks was presented to 20 subjects. Because one subject only completed 15 out of 20 sets, we had a total of 595 sets of parallel naturalness ratings for the different conditions. In 46 of these, the hidden reference was judged as less than completely natural; these sets were excluded from further analysis. The box plot in figure 1 illustrates the distribution of the remaining naturalness scores, aggregated across all subjects and prompts.

The distributions in figure 1 are broad due to variability between stimuli and listeners. Fortunately, the MUSHRA design permits the use of pairwise t-tests to identify significant differences in mean naturalness between conditions (assuming differences follow a normal distribution). Given the large number of condition pairs to compare (66 in total), it is necessary to apply a Bonferroni correction. Even so, all condition pairs were found to differ significantly in mean naturalness at a 1% level, except for (VU, V), (SF, SI), (L1, H1), (L2, H2), and (SI, M).

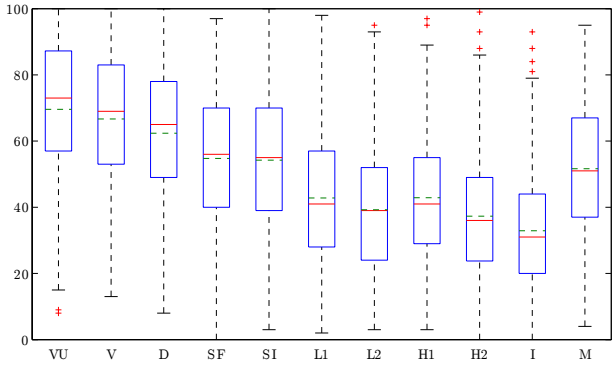

Figure 1: *Box plot of MUSHRA naturalness ratings for the different conditions, aggregated over all subjects and prompts. Solid red centre lines are medians. Dashed horizontal green lines are means. Box edges identify 25 and 75% percent quantiles. Dashed vertical whiskers generally extend to the limits of the distribution, otherwise outliers are plotted as red crosses. Condition N is omitted as it was always rated at 100.*

5.3. Discussion of the naturalness evaluation

Figure 1 shows that the conditions group into four clusters: natural speech (N), vocoded (VU, V, and D), stream independence (SF and SI), and MCEP independence (L1, L2, H1, H2, and I). The assumption that speech can be vocoded (VU) drops mean naturalness more than 30 points. The source-filter conditional independence assumption (SF) induces a notable 7.6 point mean naturalness loss from condition D. All mel cepstral conditional independence assumptions also decrease mean naturalness substantially, with losses ranging between 12 (L1 and H1) and 22 points (I) compared to condition SF. These numbers show that neither of the two independence assumption classes we investigated are adequate for generating natural sampled speech.

Within clusters, we notice other tendencies. Despite apparently removing artefacts, smoothing parameter trajectories actually introduced a small loss in naturalness (V compared to VU; significant at 5% level). DTW (D) also had a slight negative impact: even $\mathcal{M}_{\mathrm{D}}$ is not as natural as vocoded speech (VU). The closeness of SF and SI implies that, once source and filter streams are assumed independent, introducing independence between LF0 and BAPs makes no difference. (L1, H1) do not differ significantly in naturalness, and likewise for (L2, H2). That L1 and H1 are similar in naturalness suggests that, while interdependencies matter for all MCEPs, dependencies among lower-order coefficients (0–12) are relatively more important for naturalness than those among higher-order ones.

Parameter generation by averaging (M) sounds the same regardless of the model used, and is close to SF and SI in naturalness. This is better than sampling from models making further assumptions, consistent with the current preference for MLPG over sampling given the relatively simplistic models used.

Mean-based parameter generation for model $\mathcal{M}_{\mathrm{D}}$ introduces a drop of 11 points compared to sampling from the same model (M vs. D), implying that the mean of the true speech distribution in our parameter domain has lower naturalness than speech samples and calling into question mean-based parameter generation methods. Interestingly, perceived naturalness can often be improved by augmenting MLPG with *global variance modelling* (GV) [22] or *postfiltering* [23]; in the case of GV, this is actually equivalent to mean-based parameter generation with an exaggerated, less accurate acoustic model [24]. These methods are however unlikely to restore missing spectral detail.

5.4. Discrimination experiment

To complement the naturalness ratings, our second experiment explored the perceptual dimensions that listeners use to discriminate between stimuli from different conditions. 20 native speakers of English were presented with pairs of stimuli and asked whether the naturalness within each pair differed or not. The fraction of times two distinct conditions are judged as different reflects the perceptual distance between them. Random pairs of distinct conditions were presented; each subject provided 132 ratings and all 66 distinct pairs were compared 40 times. To prevent the task from being too easy (such that all conditions are judged as different), the six stimuli within any three consecutive pairs were always based on six differing prompts. Responses were accumulated in a 12×12 matrix indicating the fraction of times that each pair of distinct conditions was judged as different, with values ranging from 0.3 to 0.85. Off-diagonal values can be interpreted as a monotonic transformation of the pairwise perceptual distances between conditions.

Using *ordinal multidimensional scaling* (MDS) [25], points representing each condition can be embedded in a low-dimensional space such that the Euclidean distances between them approximates their perceptual distances. This is a widely used method for visualising discrimination data. Figure 2 shows a two-dimensional MDS analysis of or our data, computed using the `mdscale` command in the Matlab statistics toolbox with default initialisation settings.

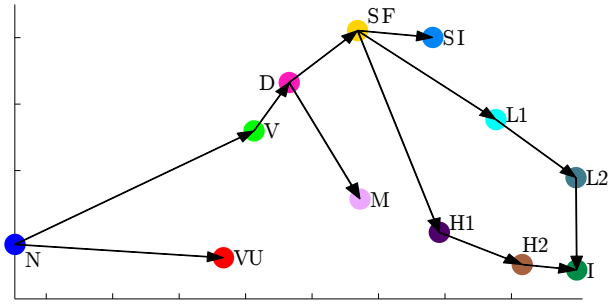

Figure 2: *Ordinal MDS visualisation of approximate perceptual distances between conditions. Each circle represents a condition; arrows illustrate the associated hierarchy of assumptions.*

5.5. Discussion of the discrimination experiment

Although the MDS plot is only an approximation of a high-dimensional perceptual space, the results are consistent with the findings from the MUSHRA test. In particular, the point configuration nicely mirrors the hierarchy of assumptions among our conditions, as made clear by the superimposed graph structure: Each cumulative assumption (arrow) takes us further from natural speech; the longest arrows are associated with the most detrimental steps, such as vocoding or making MCEP independence assumptions. We also observe other tendencies that the one-dimensional MUSHRA methodology cannot distinguish: for instance, the vocoded conditions V and VU, while similar in naturalness, are seen to be perceptually far apart.

6. Conclusions and future work

Through subjective listening tests, we have identified fundamental naturalness limitations imposed by assumptions used in statistical parametric speech synthesis. In particular, assuming source and filter parameter streams to be conditionally independent, or assuming full or partial independence among mel cepstral coefficients, both significantly limit the naturalness of speech generated by sampling. The mean trajectories of near-perfect acoustic models, computed by averaging over the database, suffer from naturalness deficiencies regardless of the independence assumptions considered. This limits the performance of mean-based parameter generation methods such as no-GV MLPG with trajectory HMMs. However, it is possible that greater perceived naturalness can be attained by using other parameter generation methods or by applying standard approaches in better speech parameter representations.

The methodology presented in this paper can straightforwardly be extended to examine the perceptual consequences of additional modelling assumptions, all the way to a full TTS system. This is in progress using an expanded repetitions database.

Acknowledgements: The research leading to these results was partly funded by EPSRC Programme Grant no. EP/I031022/1 (Natural Speech Technology) and by the European Community 7th Framework Programme Marie Curie INSPIRE ITN. We are indebted to Dr. Oliver Watts for suggesting that we use our material to investigate stream independence assumptions.

7. References

[1] S. King and V. Karaiskos, "The Blizzard Challenge 2012," in *Proc. Blizzard Chall. Workshop*, 2012.

[2] M. Shannon, H. Zen, and W. Byrne, "The effect of using normalized models in statistical speech synthesis," in *Proc. Interspeech*, 2011.

[3] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*, vol. 3, 2000, pp. 1315–1318.

[4] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *Proc. SSW8*, 2013, pp. 155–160.

[5] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE T. Speech Audi. P.*, vol. 7, no. 3, pp. 272–281, 1999.

[6] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. SSW6*, 2007, pp. 294–299.

[7] T. Merritt and S. King, "Investigating the shortcomings of HMM synthesis," in *Proc. SSW8*, 2013, pp. 185–190.

[8] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Upper Saddle River, NJ: Prentice Hall, 2001, p. 12.

[9] T. Shinozaki and S. Furui, "An assessment of automatic recognition techniques for spontaneous speech in comparison with human performance," in *Proc. SSPR*, 2003.

[10] A. Juneja, "A comparison of automatic and human speech recognition in null grammar," *J. Acoust. Soc. Am.*, vol. 131, no. 3, pp. EL256–EL261, 2012.

[11] D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch," in *Proc. ICSLP*, 1998.

[12] D. Gillick, L. Gillick, and S. Wegmann, "Don't multiply lightly: quantifying problems with the acoustic model assumptions in speech recognition," in *Proc. ASRU*, 2011, pp. 71–76.

[13] S. King, "An introduction to statistical parametric speech synthesis," *Sadhana*, vol. 36, no. 5, pp. 837–852, 2011.

[14] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[15] G. Fant, *Acoustic Theory of Speech Production: With Calculations based on X-Ray Studies of Russian Articulations*, 2nd ed. The Hague, The Netherlands: Mouton & Co., 1970.

[16] H. Zen, K. Tokuda, and T. Kitamura, "Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences," *Comput. Speech Lang.*, vol. 21, no. 1, pp. 153–173, 2007.

[17] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statistical parametric speech synthesis," *IEEE T. Audio Speech*, vol. 21, no. 3, pp. 587–597, 2013.

[18] E. H. Rothauser, W. D. Chapman, N. Guttman, K. S. Nordby, H. R. Silbiger, G. E. Urbanek, and M. Weinstock, "IEEE recommended practice for speech quality measurements," *IEEE T. Acoust. Speech*, vol. 17, no. 3, pp. 225–246, 1969.

[19] *Objective measurement of active speech level*, ITU Recommendation ITU-T P.56, International Telecommunication Union, Telecommunication Standardization Sector, Geneva, Switzerland, March 2011.

[20] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349–353, 2006.

[21] *Method for the subjective assessment of intermediate quality level of coding systems*, ITU Recommendation ITU-R BS.1534-1, International Telecommunication Union Radiocommunication Assembly, Geneva, Switzerland, March 2003.

[22] T. Tomoki and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.

[23] H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. Interspeech*, 2012.

[24] M. Shannon and W. Byrne, "Fast, low-artifact speech synthesis considering global variance," in *Proc. ICASSP*, 2013, pp. 7869–7873.

[25] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, 2nd ed. New York, NY: Springer, 2005.