A DEEP LEARNING BASED ANALYSIS-SYNTHESIS FRAMEWORK FOR UNISON SINGING

Pritish Chandna[1]   **Helena Cuesta**[1]   **Emilia Gómez**[2,1]

[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona
[2] European Commission, Joint Research Centre, Seville

`{pritish.chandna, helena.cuesta, emilia.gomez}@upf.edu`

ABSTRACT

Unison singing is the name given to an ensemble of singers simultaneously singing the same melody and lyrics. While each individual singer in a unison sings the same principle melody, there are slight timing and pitch deviations between the singers, which, along with the ensemble of timbres, give the listener a perceived sense of "unison". In this paper, we present a study of unison singing in the context of choirs; utilising some recently proposed deep-learning based methodologies, we analyse the fundamental frequency (F0) distribution of the individual singers in recordings of unison mixtures. Based on the analysis, we propose a system for synthesising a unison signal from an *a cappella* input and a single voice prototype representative of a unison mixture. We use subjective listening tests to evaluate perceptual factors of our proposed system for synthesis, including quality, adherence to the melody as well the degree of perceived unison.

1. INTRODUCTION

Throughout history, singing has been an important cultural activity for humans, serving for propagation of beliefs and ideas amongst the masses as well as for social entertainment. The social aspect led to gatherings of people singing in a group, which evolved into polyphonic ensemble singing with multiple voices singing counterpoint melodies in complex harmonies. A group of people singing in such an ensemble is commonly termed as a choir and the focus of our study is on one setting of such choirs consisting of four voices known as Soprano, Alto, Tenor and Bass (SATB). Each voice within an SATB ensemble has its own function and melodic range in the whole. SATB is one of the most widely studied, documented, and practiced forms of choirs with numerous dedicated conservatories across Europe, highlighting the cultural importance of the art form. Within each of the SATB voices, it is common to have multiple singers of similar

vocal range singing the same melody simultaneously, in a form known as unison singing. While all the singers in a unison sing the same melody, it is impossible for a group of two or more people to perfectly synchronize and sing the exact same pitch line. Each singer has their own natural micro-deviations, both in terms of timing and pitch, from the prescribed score and their own distinct timbre. The combination of micro-deviations and the ensemble of timbres leads to the perception of unison, wherein several singers are perceived to be singing a single pitch contour [1], and is the main focus of our study.

Pitch and fundamental frequency (F0) are related but not equivalent terms. While the F0 generally refers to the physical frequency of vibrations of the vocal folds for a singing voice signal, pitch refers to an abstract perceptual concept which has been found to be closely correlated to the F0. Frequency is usually measured in Hertz, representing the number of cycles of a periodic signal per second, whereas pitch is described in terms of perceptual units like *cents*. The cent is a unit defined on a logarithmic scale, as a measure of the ratio between the frequency in Hertz and a base frequency, commonly chosen to be $440\,\mathrm{Hz}$, as shown in Equation 1.

$$f_{cents} = 1200 \cdot log_2 \frac{f_{hertz}}{440} \qquad (1)$$

Thus defined, the cent is correlated to the perceptually relevant musical unit of an equally tempered semitone. Specifically, one semitone spans $100\,\mathrm{cents}$. Examined individually, the pitch of the singers in a unison can be represented by the F0 of each individual singer's vocal signal, this can be tested by synthesising a time-varying sinusoid with the frequency of the signal. However, when the individual signals are added the resultant pitch is not merely the sum of F0 value, and the methodology of synthesising a sum of sinusoids as with single singers fails to produce the same perceptual result due to physical phenomena such as beating, among others. Past studies have utilized artificial unison mixes created by the use of a vowel only singing voice synthesizer to study the perception of a single pitch contour in a unison [1]. Other areas of past research related to unison singing include single voice to unison synthesis models, based on creating voice clones with variations in the input [2, 3]. Fuelled by the deep learning revolution, singing voice synthesizers have evolved over the

last few years, allowing us to take a step further both into exploring the perception of unison and into effective solo voice to unison synthesis. We build on the work done by Ternström [1] by leveraging recently proposed synthesis methodologies to synthesize a single voice prototype representing the melodic and linguistic content of a unison mixture. This allows use to further test the hypothesis of a single F0 contour representative of the perceived pitch of a unison via subjective listening tests. We also verify the author's findings by analysing a set of real recordings of unison singing. In addition, we propose a methodology combining previous research and recently developed techniques to synthesize a unison mixture from a single voice input. We follow the basic methodology used by Schnell et al. [2] to create voice clones with variations in three aspects; pitch, timing and timbre, and use perceptual evaluation tests to evaluate the effect of each of these parameters on the perception of the sense of a unison.

The rest of the paper is structured as follows. Section 2 discusses previous works pertinent to our study. We then present the analysis of the choir recordings in Section 3, including a description of the dataset of choir recordings, the methodology used for the analysis and the results of the analysis. The synthesis methodology we use for synthesizing voice clones and the single voice prototype of the unison mixture is described in Section 4. Section 4.3 presents the perceptual evaluation methodology used and the results of the perceptual tests. Finally, we present a discussion on our findings in the analysis of the choir and the perceptual evaluation of the synthesis in Section 6.

2. RELATED WORK

We divide the description of related works into three sections: past studies into the analysis of the perception of unison, previous works on synthesising unison mixtures from choirs and recently proposed deep-learning based methodologies what we will use for analysis and synthesis.

2.1 Analysis Of Unison Perception

The perception of pitch dispersion has previously been studied in [1], wherein the author used synthesized singing voice stimuli to investigate the preferences of expert listeners in unisons. In the study, pitch dispersion is defined as the bandwidth of the fundamental frequency and its harmonic partials across individual singers in a unison. It is suggested that this dispersion is related to the *flutter*— small variations in F0 that are too fast to be perceived as pitch variations. The concept of pitch *scatter* is presented in the study as the standard deviation over voices in the mean F0: the average F0 computed over the duration of each tone of a song. The study concludes that a *scatter* of 0 cents–5 cents was preferred by the participants while a *scatter* of 5 cents–14 cents was seen as the limit of tolerance before dissonance was reported. In addition, the author also highlights several differences between solo and ensemble singing. For instance, a single performer produces tones with well-defined properties: pitch, timing,

loudness, timbre, while an ensemble of performers produces sounds with statistical distributions of each of these properties.

A similar method for modelling *scatter* in choir sections was presented by Cuesta et al. [4] using small windows to compute the standard deviation between individual F0s in the unison. This study used real recordings of choral singing instead of synthesised stimuli, presenting a mathematical model for dispersion rather than a perceptual evaluation. For the dataset used in the research, F0 (or pitch) dispersion was found to be in the range of 20 cents–30 cents for all SATB voice sections, being slightly larger in the Bass.

Another recent study focused on the analysis of F0 in vocal music is work by Weiss et al. [5], where the authors proposed an approach to measure intonation quality of choir recordings. They create an ideal 12-tone equal temperament grid, and then calculate the deviation of each F0 and its partials to their theoretical position in the grid. The overall deviation is computed as a weighted sum of each partial's deviation. This method enables the analysis of the overall intonation of a full choir recording, but does not account for the deviations within sections of the choir.

2.2 Unison Synthesis

Signal processing techniques have previously been utilised to synthesize choir unison by adding several clones of a monophonic *a cappella* signal with uncorrelated pitch, timing, and timbre deviations. Most particularly, Pitch Synchronous Overlap Add (PSOLA) methods [2] have been exploited as an analysis-synthesis framework to decompose the vocal signal into a set of constituent waveforms representing successive pitch periods of the signal. Pitch and timing deviations are added to the vocal signal using time stretching and pitch shifting techniques to create voice clones, which are combined to form the output unison signal.

Other proposed methodologies for creating a unison output from an *a cappella* signal include morphing the spectral and pitch components of the vowels of the input signal as in [3]. The methodology's effectiveness is constrained to low tempo inputs . Random modulation of beating partials to create a choral effect [6] has also been used.

2.3 Deep Learning For Analysis and Synthesis

For our work, we build on the work done in [1] and [4], modelling the perceptual pitch contour of a unison mixture as a single F0 contour. To this end, we use a recently proposed Convolutional Representation for Pitch Estimation (CREPE) methodology [7] for extracting F0 contours from real world recordings of individual singers in a choir setting as well the F0 contour of unison mixture created by combining the individual voices. This methodology uses a series of convolutional operations on the waveform of the input signal and outputs a probability distribution over a discrete representation of the underlying F0 contour of the signal across a series of time-frames.

To synthesize the single voice prototype representing a unison mixture output and the voice clones for creating a unison mixture from a single voice input, we adapt the methodology proposed by Chandna et al. [8], which allows for the re-synthesis of a solo single voice from a musical mixture input via the underlying linguistic features. This methodology builds on the idea of re-synthesizing a vocal signal from a musical mixture by estimating the parameters of a vocoder synthesizer [9] and uses an encoder built of a bank of bi-directional long short-term memory (LSTM) recurrent neural networks (RNNs) to estimate a continuous representation of the underlying linguistic features present in the input mixture signal. The continuous representation is singer-independent and language agnostic, and was initially proposed for zero-shot voice conversion via an autoencoder network [10]. The linguistic features can then be used to generate the spectral envelope of the vocal signal in the mixture, providing the singer identity. The authors of [8] proposed two decoders for this process, a Singer Dependent Network (SDN) which takes the singer identity as a one-hot vector, and a Singer Independent Network (SIN) which intrinsically learns the singer identity from the given input. The spectral envelope is then combined with the F0, extracted via an external algorithm, to synthesize the vocal signal. While the original framework was proposed for extracting a singing voice from a pop/rock musical mixture, we adapt the SDN network for synthesising a unison mixture from an *a cappella* input and the SIN network for synthesizing an *a cappella* singing voice from a unison mixture. The adaptations we apply are described in Section 4. The SIN and SDN models [8] were trained on a proprietary dataset with 12 hours of data comprising 205 songs by 45 male and female singers, and we have obtained a copy of the trained model with permission from the relevant authorities for our study.

3. ANALYSIS OF CHOIR RECORDINGS

We analyse the variations between individual singers in a unison in terms of variance in pitch and timing. Below, we present the dataset that we use in our analysis, followed by the methodology used for analysis, and finally the results of our analysis.

3.1 Datasets

We analyse the Choral Singing Dataset (CSD) [4], which includes monophonic recordings of 3 choral pieces: *Niño Dios d'Amor Herido*, written by Francisco Guerrero, *Locus Iste*, written by Anton Bruckner, and *El Rossinyol*, a Catalan popular song. There are 16 different singers for each song with 4 singers for each of the four parts; Soprano, Alto, Tenor and Bass. The dataset also includes manually corrected F0 annotations for each track.

3.2 Analysis Methodology

To analyse inter-singer variance in **pitch**, the first step is the extraction of an F0 contour from a unison mixture of singers. We aim to study the behavior of a monophonic

F0 extractor in such cases, assuming that we have a sufficiently balanced unison performance, where the contribution of each singer is similar in terms of volume and energy. To this end, we use CREPE [7] to extract the fundamental frequency of the unison mixture created by summing and normalizing all corresponding individual singers in each vocal part of the recordings. This is termed as $EstF0_U$.

We then measure the resemblance of the estimated $EstF0_U$ to each of the manually annotated F0 tracks and to the mean $F0_m$ [1]. We use standard evaluation metrics for melody extraction including *Raw Pitch Accuracy (RPA), Overall Accuracy (OA), Voicing Recall (VR)* and *Voicing False Alarm (VFA)* between the $EstF0_U$, the average the mean $F0_m$, and each individual singer curve, $GTF0_i$ [2].

Once we have verified the accuracy of the extraction system, we build a statistical model for the individual contours in the unison, as suggested by [1]. In our model, the framewise F0 of an individual singer, $F0_i$, can be represented as a distribution of values around the mean $F0_m$ with a deviation of F_{devi}, as shown in Equation 2

$$F0_i = F0_M + F_{devi} \qquad (2)$$

This equation also allows us to define the $F0_{i+1}$ of a singer in terms of the $F0_i$ of another singer in the unison as

$$
\begin{aligned}
F0_{i+1} &= F0_m + F_{devi+1} \\
F0_{i+1} - F0_i &= F_{devi+1} - F_{devi} \\
F0_{i+1} &= F0_i + F_{devi+1} - F_{devi} \\
F0_{i+1} &= F0_i + \Delta F0_s
\end{aligned}
\qquad (3)
$$

Where we define $\Delta F0_s$ as the inter-singer deviation, represented by Equation 4. For each pair of singers in the unison, we compute the frame-wise difference between the corresponding F0 contours in cents. For this calculation, only frames with positive F0 values, also known as *voiced* frames, were considered. We average these inter-singer deviations across time and songs, and obtain a single value for each group, i.e., SATB.

$$\Delta F0_s = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \left| F0_i - F0_j \right|}{\binom{n}{2}} \qquad (4)$$

where the sub-index s indicates the choir section, $s \in [S, A, T, B]$, and n is the number of singers. In our use case, $n = 4$.

To study **timing** deviations, we focus in the transitions from *voiced* to *unvoiced*, and vice-versa—where singers are not entirely in sync. We call these regions *transition regions*, where some of the singers in the mixture are voiced

---

[1] Note that the average F0 value has to be adjusted for timing differences between the individual singers. To this end, we define the average to be zero (unvoiced frame) if and only if all individual values for that frame are zero. for all other cases, the average is calculated only accounting for the non-zero values.

[2] We use the `mir_eval` library [11] for this evaluation, and we use a pitch tolerance of $30\,\text{cents}$
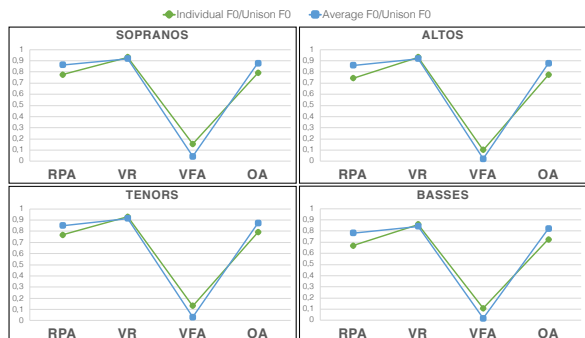
Figure 1. Resemblance of the estimated unison $EstF0_U$ estimation to each individual $GTF0_i$ contour (green) and the average (blue) using pitch evaluation metrics averaged across each choir section.
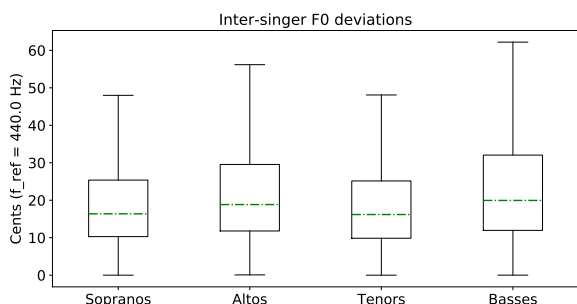
Figure 2. Inter-singer deviations in cents averaged across the whole dataset for each choir section. Deviations are calculated using Equation 4.

and others are unvoiced, with a positive or zero F0. We measure the length of all the transition regions in every unison from the CSD, and then average across choir sections.

3.3 Analysis Results

A summary of the results of the comparison between the fundamental frequency extracted by CREPE, $EstF0_U$, and the manually corrected fundamental frequency, $GTF0_i$ is illustrated in Figure 1, along with a comparison with the mean, $F0_m$. We observe that all sections follow the same pattern with similar metric values, and the unison F0 estimated by CREPE, $EstF0_U$, is closer to the average $F0_m$,than to the individual contours. In addition, all metrics improve when we compare the average F0 curve to the extracted F0 contour from the unison: RPA, VR and OA are higher in the blue plots, while VFA is lower. We can thus use the pitch estimated by CREPE, $EstF0_U$, as a representative of the mean single pitch contour perceived in a unison mixture [1].

The calculated $\Delta F0_s$ is shown in Figure 2. We observe an inter-singer deviation in the range of 0 cents–50 cents, with a mean of around 20 cents. This value, representing the inter-singer deviation in the unison mixtures, is comparable to the pitch dispersion studied by Cuesta et al. [4]. While the methodology for modelling is different, these re-

Section	Average Timing Deviation $\pm$ Standard Deviation
Soprano	0.134 ± 0.039 sec
Alto	0.093 ± 0.0024 sec
Tenor	0.100 ± 0.021 sec
Bass	0.124 ± 0.021 sec

Table 1. Timing deviations averaged across the CSD. These values measure the time span in which all singers in the unison transition from voiced to unvoiced, and vice-versa, averaged across all transitions in each song.

sults are in accordance with their reported per-section pitch dispersion: larger in the bass section, smaller in the sopranos, and very similar for altos and tenors.

Table 1 shows the results of the timing analysis. We observe an average timing deviation of 0.1 s between the voices in the unison for all parts of the choir.

4. SYNTHESIS METHODOLOGY

We present two synthesis models, Solo To Unison (STU) to synthesize voice clones for creating a unison mixture from a single voice input, and Unison to Solo (UTS) for synthesizing single voice prototype representing the melodic and linguistic content of a unison mixture input[3]. Similar to the work presented by Schnell et al [2], we decompose the input signal into the F0, harmonic spectral envelope, and aperiodicity envelope. However, instead of using Pitch Synchronous Overlap Add (PSOLA) methods, we utilize the WORLD vocoder [12], which has been shown to be an effective vocoding system for singing voice synthesis [13, 14, 14]. Similar to [14], we use truncated frequency warping in the cepstral domain [15] to reduce the dimensions of the harmonic components from 1024 to 60 log Mel-Frequency Spectral Coefficients (MFSCs) with an all-pole filter with warping coefficient $\alpha = 0.45$. In addition, we use bandwise aperiodic analysis to reduce the dimensionality of aperiodic features to 4. For the rest of this paper, we refer to these 64 features together as the spectral envelope.

4.1 Unison to Solo (UTS)

As shown in Figure 3, we first perform a short-time Fourier transform (STFT) to extract a spectrogram from the input waveform. The magnitude part of the spectrogram is passed through the encoder proposed in [8] to extract a continuous representation of the linguistic features present in the unison mixture input. The linguistic features are decoded via the SIN network [8] to generate the spectral envelope for vocal synthesis. This envelope is combined

---

[3] Audio examples are provided as complementary material at `https://pc2752.github.io/unison_analysis_synthesis_examples/`and the source code with the trained models are available at `https://github.com/MTG/content_choral_separation`

with the pitch contour output from CREPE [7], $F0_U$ to synthesise the single voice prototype representing the unison mixture input.

4.2 Solo to Unison (STU)

The analysis part of the STU case follows a similar methodology, as we extract the linguistic features and the $EstF0_i$ contour from the input *a cappella* voice signal. To create voice clones with pitch and timing deviations, we add randomly sampled noise from a normal distribution with a mean of 0 and a variable standard deviation, termed as std. This represents the inter-singer deviation, $\Delta F0_s$, and allows us to model the $F0_{i+1}$ of the clone as per Equation 3. Timing deviations are added by shifting the voiced portions of the input signal or the portions between tow blocks of silence of more than 80 ms by a variable amount, randomly sampled from a normal distribution of mean 0 and standard deviation ts. The values of std and ts are based on our analysis of the Choral Singing Dataset presented in Section 3.3

Finally, for variations in timbre, we generate the spectral envelope of a variable number singers, ns, of the same gender as the input using the SDN network proposed in [8]. This is based on our analysis presented in Section 3.3. There was no overlap between singers in the set used for training the synthesis model and the singers in the Choral Singing Dataset used for evaluation. The various voice clones are added together and normalized in amplitude to produce the final unison output. We evaluate various combinations of std, ts, and ns on their impact of the perception of unison.

4.3 Perceptual Evaluation Methodology

We used subjective listening tests with low and high anchors, as modified versions of the MUSHRA-methodology [16] to evaluate subjective criteria of the synthesis produced by our analysis synthesis framework.

While there are several aspects that could be evaluated, we focused on three keys aspects: adherence to melody, perception of unison, and audio quality. For each aspect, the participants were presented with 4 questions, one for each part of the SATB choir, and were asked to rate the test cases in the question on a continuous scale of 1–5 with respect to a presented reference. The test case and references provided pertained to the the same section of the song and were between 7 s–10 s each. The parameters used for these tests are described below for each aspect.

4.3.1 Adherence to melody and lyrics

For this aspect, we wanted to see the similarity of the perceived pitch contour of the output for both the UTS and STU cases to that of a ground truth unison mixture. To this end, the reference provided to the participant was a ground truth unison sample made by summing the corresponding four individual singers of a part to form a unison mixture. This reference is referred to as *REFU*. The participants were asked to rate test samples which included the single voice prototype of the unison as output by the UTS

system, referred to as *UTS*. In addition, we evaluated the output of STU with a pitch deviation with parameter std set to 50 cents, the acceptable limit of pitch deviations, as shown by our analysis in Section 3.3 and suggested by [1]. Four singers were used for generating this test case, with parameter ns set to 4, and it is referred to as *STU_PS*. We also evaluated the output of the UTS system with both pitch and timing deviations with parameter ts set to 40 ms. While our analysis in Section 3.3 suggests that higher values of ts could have been used, we found that increasing the value beyond 40 ms leads to a unacceptable level of degradation in output quality. We refer to this test case as *STU_PTS*. We also provided a lower anchor of a sample of the same length from another vocal part.

4.3.2 Perception of unison

Unison is a loosely defined perceptual aspect, the cognition of which we aim to study here. For this, we provide a reference of a ground truth unison sample created in the same manner as described above, *REFU*. Given this, participants were asked to rate outputs from the STU system based on their similarity to the reference in terms of the perception of unison. In addition to the *STU_PTS* and *STU_PS* cases with pitch, timing and timbre variance, we also tested the case for just timing and singer variation, referred to as *STU_TS* and a case with just pitch and timing deviations, referred to as *STU_PT*, timbral changes were not done for the voice clones used for creating this test case. The *a cappella* sample of a single singer singing the same example as the reference was provided as a lower anchor.

4.3.3 Audio Quality

Audio quality is another subjective measure that is well defined in literature but not easily understood by non-expert participants. For the evaluation, we set an upper limit of audio quality to the resynthesis of a single voice recording with the WORLD vocoder *REFS* and a lower limit to the resynthesis of a unison mixture with the same *RESSYN-THU*. The test cases provided to the participants were the same as those provided for the adherence to melody case, except that the lower anchor was changed.

4.4 Perceptual Evaluation Results

There were 17 participants in our evaluation, of which 10 had prior musical training. To account for inter-participant variance in subjective evaluation, the opinion score for each question was normalized over ratings for the reference and the lower anchor before calculating the mean opinion scores (MOS) and the standard deviations in opinion scores, presented in Table 2.

The subjective nature of the perceptual aspects evaluated must be taken into account for the evaluation and the mean opinion scores are indicative of preferences rather than absolute measures of quantity. It can be observed that the perceived adherence to melody for the prototypical *a cappella* voice synthesized by the UTS model has higher preference than the STU models, although a high variance
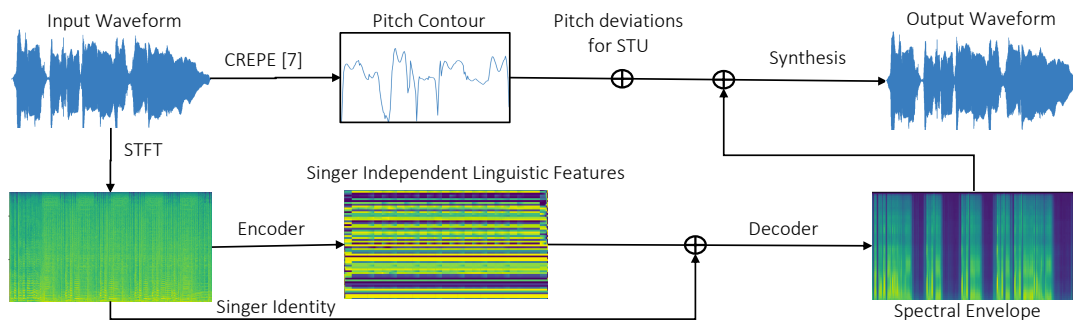
Figure 3. The synthesis framework. The magnitude part of the spectrogram of the input is passed through the encoder from [8] to extract linguistic content. For the UTS case, this is passed to the decoder along with a learned embedding to the SIN decoder [8] to generate the spectral envelope. For STU, the linguistic features are passed to the SDN decoder along with a one-hot vector of the same gender as the input. F0 is extracted from the input waveform via CREPE [7]. Both the envelope and the F0 are used to synthesize the output voice. For the STU case, timing deviations are further added before summing with the input and normalizing.

Test Case	Adherence To Melody	Unison Perception	Audio Quality
UTS	3.6 ± 0.93		2.1 ± 0.65
STU_PS	3.3 ± 0.83	2.6 ± 0.85	2.8 ± 0.45
STU_PTS	2.9 ± 1.14	3.2 ± 0.96	3.1 ± 0.63
STU_TS		2.3 ± 1.11	
STU_PT		3.0 ± 1.23	

Table 2. Mean Opinion Score (MOS) $\pm$ Standard Deviation for the perceptual listening tests across the test cases provided. The models shown corresdond to the Unison to Solo (UTS) model, the Solo to Unison with pitch, timing and singer variations, indicated by the addition of the letters P,T and S as suffixes to the abbreviation, respectively. The scores for each question were normalized by the responses to the upper and lower limits for the responses defined in section 4.3.

is observed in the ratings for the same. The unison perception evaluation shows that the variations in either timing or pitch alone are not as preferred as variations in both aspects together. Timbre variations do not have as significant an effect on perception of unison as variances in timing and pitch. The evaluation of audio quality shows room for improvement in the synthesis of the voice signals. This can partly be attributed to the use of the WORLD vocoder [12] and we believe that this can be improved on in the future using recently proposed neural synthesis techniques.

5. CONCLUSIONS

We have presented an analysis of the Choral Singing Dataset, building on the work presented in [1]. In accordance with the analysis done by [4], we observe deviation in the range of $0\,\mathrm{cents}$–$50\,\mathrm{cents}$ between the F0 contours of the individual singers in the unison mixtures in the dataset. We further note an timing deviation of $0.1\,\mathrm{s}$ between singers in unison in the dataset.

We then used this analysis along with recently proposed deep-learning based methodologies to present a synthesis

system for a unison mixture from a single voice input and a single voice prototype synthesis representing the melodic and linguistic content of a unison mixture input. Based on these systems, we were able to conduct a perceptual evaluation of the unison, further supporting the claim of [1] that the a mixture of different voices singing in unison is perceived to have a single pitch. In addition, we found that pitch and timing deviations together are important for the perception of the unison, and that variations in either aspect alone is insufficient for such. However, timbre variations were not found to be as relevant.

We present this work as the first step into the analysis of an under-explored research area, hoping to fuel further discussion on the topic. While interesting from an academic standpoint, the systems we present also have several commercial applications such as creating a unison choral effect to be used in music production as well as for transposition and transcription, in conjunction with the work presented in [17]. We also plan to incorporate the presented work with [18], for complete source separation for choral recordings.

6. ACKNOWLEDGEMENTS

7. REFERENCES

[1] S. Ternström, "Perceptual evaluations of voice scatter in unison choir sounds," *STL-Quarterly Progress and Status Report*, vol. 32, pp. 041–049, 1991.

[2] N. Schnell, G. Peeters, S. Lemouton, P. Manoury, and X. Rodet, "Synthesizing a choir in real-time using pitch synchronous overlap add (psola)." in *ICMC*, 2000.

[3] J. Bonada, "Voice solo to unison choir transformation," in *Proceedings of the 118th Audio Engineering Society Convention*, Barcelona, Spain, May 2005.

[4] H. Cuesta, E. Gómez, A. Martorell, and F. Loáiciga, "Analysis of intonation in unison choir singing," in *Proceedings of the International Conference of Music Perception and Cognition (ICMPC)*, Graz, Austria, 2018, pp. 125–130.

[5] C. Weiss, S. J. Schelcht, S. Rosenzweig, and M. Müller, "Towards measuring intonation quality of choir recordings: A case study on bruckner's locus iste." in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, Delft, The Netherlands, 2019, pp. 276–283.

[6] M. B. Dolson, "A tracking phase vocoder and its use in the analysis of ensemble sounds," Ph.D. dissertation, California Institute of Technology, 1983.

[7] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Canada, 2018, pp. 161–165.

[8] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "Content based singing voice extraction from a musical mixture," in *Proceedings of the 45th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2020, pp. 781–785.

[9] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "A vocoder based method for singing voice extraction," in *Proceedings of the 44th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2019)*, IEEE. Brighton, UK: IEEE, 2019. [Online]. Available: https://arxiv.org/abs/1903.07554

[10] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," in *International Conference on Machine Learning*, 2019, pp. 5210–5219.

[11] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. Ellis, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[12] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[13] P. Chandna, M. Blaauw, J. Bonada, and E. Gómez, "WGANsing: A multi-voice singing voice synthesizer based on the wasserstein-gan," in *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 2019, pp. 1–5.

[14] M. Blaauw and J. Bonada, "A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs," *Applied Sciences*, vol. 7, no. 1313, 12/2017 2017.

[15] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in *3rd International Conference on Spoken Language Processing (ICSLP)*, 1994.

[16] C. Völker, T. Bisitz, R. Huber, B. Kollmeier, and S. M. Ernst, "Modifications of the multi stimulus test with hidden reference and anchor (mushra) for use in audiology," *International journal of audiology*, vol. 57, no. sup3, pp. S92–S104, 2018.

[17] H. Cuesta, B. McFee, and E. Gómez, "Multiple F0 Estimation in Vocal Ensembles using Convolutional Neural Networks," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montreal, Canada (Virtual), 2020.

[18] D. Petermann, P. Chandna, H. Cuesta, J. Bonada, and E. Gómez, "Deep Learning Based Source Separation Applied To Choir Ensembles," in *Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR)*, Montreal, Canada (Virtual), 2020.