THE USE OF VOICE SOURCE FEATURES FOR SUNG SPEECH RECOGNITION

Gerardo Roa Dabike, Jon Barker

Department of Computer Science, University of Sheffield, UK

ABSTRACT

In this paper, we ask whether vocal source features (pitch, shimmer, jitter, etc) can improve the performance of automatic *sung* speech recognition, arguing that conclusions previously drawn from spoken speech studies may not be valid in the sung speech domain. We first use a parallel singing/speaking corpus (NUS-48E) to illustrate differences in sung vs spoken voicing characteristics including pitch range, syllables duration, vibrato, jitter and shimmer. We then use this analysis to inform speech recognition experiments on the sung speech DSing corpus, using a state of the art acoustic model and augmenting conventional features with various voice source parameters. Experiments are run with three standard (increasingly large) training sets, DSing1 (15.1 hours), DSing3 (44.7 hours) and DSing30 (149.1 hours). Pitch combined with degree of voicing produces a significant decrease in WER from 38.1% to 36.7% when training with DSing1 however smaller decreases in WER observed when training with the larger more varied DSing3 and DSing30 sets were not seen to be statistically significant. Voicing *quality* characteristics did not improve recognition performance although analysis suggests that they do contribute to an improved discrimination between voiced/unvoiced phoneme pairs.

Index Terms— Sung speech, voice source, speech recognition

1. INTRODUCTION

Automatic sung speech recognition is attracting increased research attention, driven in part by the release of large sung speech datasets, e.g., in 2020, the Music Information Retrieval Exchange (MIREX) ran its first Lyrics Transcription task[1]. Lyric transcription is an interesting problem in its own right, but equally importantly, adapting speech technology to sung speech may also provide insight into how to extend to other types of atypical speech (e.g., dysarthric speech). Further, poorly-intelligible sung speech acts as a useful stress-test for state-of-the-art (SOTA) acoustic modelling (AM) techniques.

Existing automatic sung speech recognition systems are typically based on successful approaches for spoken speech [1, 2, 3]. In particular, they use the same acoustic features. This has been motivated by the idea that spoken and sung speech share the same production system, and that semantic information is conveyed in the same way in both speech styles. However, there are several differences between sung and spoken speech that make the former more difficult to recognise, such as, the pitch ranges, the syllable duration, the existence of vibrato in singing and the fundamental fact that in singing, intelligibility is often less important than artistic interpretation. Since most of these differences centre around the use of the voice source, informing recognition systems with voice source information may help to improve their performances.

Traditional SOTA ASR systems designed for spoken speech, typically use Mel frequency cepstral coefficients (MFCC) or filter-

---

[1] https://www.music-ir.org/mirex/wiki/2020:Lyrics_Transcription

bank acoustic features to capture the vocal filter characteristics, and i-Vector speaker-specific representations [4] for speaker [5, 6] and environment adaptation [7]. These systems produce excellent results. However, unaccompanied sung speech ASR scenarios systems that utilised a combination of MFCCs and i-Vectors have not been achieving the same level of performance [8, 3]. This reduced performance may be partly explained by the relative lack of training data availability, the utilisation of inappropriate language models, and the (often) lower intelligibility of the signal. Additionally, sung speech possesses a higher pitch variability than spoken speech, e.g., females average 342 Hz when singing and 237 Hz speaking (Section 2). This variability means that one singer can effectively have different 'voices' (guided by the song structure), which may make it difficult to characterise the speakers adequately.

The inclusion of pitch information has been improving the performance of ASR systems in different spoken speech scenarios. In tonal languages systems, such as Mandarin [9], it has been widely used since the tone has a direct relation with the word meaning. In non-tonal languages systems, like English [10, 11, 12], the pitch-strength information helps to improve the discrimination between voiced and unvoiced sounds [10, 11]. Additionally, pitch information has been used for vocal tract normalisation (VTN), by exploiting the relationship between the pitch values and the vocal tract length, e.g., male speakers have a larger vocal tract than females and produce a lower pitch [12]. For children's ASR systems, pitch informed VTN has also been employed to train on adult speech data and testing in children data, by pitch-adaptive front-end [13, 14].

However, the role of the voice source in singing is fundamentally different from its role in speaking, meaning it has very different characteristics. First, for sung speech, the pitch range is larger, and the average pitch is higher than in spoken speech [15]. Second, the spoken speech pitch varies freely rising and falling within one syllable, with changes up to 12 semitones [16]. In contrast, the pitch in sung speech is expected to remain steady during a note, having controlled and discrete variations, with rather infrequent changes greater than two semitones [17]. Third, in singing, in order to carry the melody, speech is more heavily voiced, and this may interfere with the role of voicing as a phonetic cues. This may lead to difficulties in reliably discriminating between voiced/unvoiced phoneme pairs. Fourth, the duration of the sung vowels is often larger than in spoken speech, as needed to achieve the rhythm dictated by the musical composition and to convey the pitch assigned to the syllable. Finally, for artistic expression, singers employ a frequency modulation called vibrato, a musical effect corresponding to small and periodic variations of the pitch between 5.5 and 7.5 Hertz (Hz), with characteristics that are specific to the singer [18].

Considering these differences, the voice source features may be less useful as a phonetic cue for sung speech than for spoken speech. Alternatively, with sufficient data, voiced source features, and in particular, pitch, may be, on balance, beneficial for sung speech ASR systems, not only by improving the voiced and unvoiced sounds

ICASSP 2021

discrimination and sung vowel classification [19], but also by helping the system to normalise for different speakers or for systematic changes in phoneme quality apparent when a singer is singing in a different part of his or her vocal range. Additionally, using characterisation of the voice source *quality*, which are strongly associated with speaker identity, may also help the system by providing conditioning variables that improve speaker normalisation.

This paper aims to critically evaluate the usefulness of various types of voice source information, for improving the performance of sung speech ASR systems. This is the first study using a system tested and trained on large amounts of sung speech data. (Pitch information was previously evaluated in [19] but in a system trained using spoken data and evaluated only with male singers). Section 2 presents a data-driven analysis of some of the differences between sung and spoken speech. Section 3 details the voice source-based features chosen for use in the ASR experiments, informed by this analysis. Section 4, describes the speech recognition dataset and the baseline system. The experimental results and analysis are presented in Section 5 and conclusions in Section 6.

2. SUNG SPEECH ANALYSIS

Using the sung and spoken lyrics corpus NUS-48E [20], we examined some of the sung speech characteristics that makes it more challenging than spoken speech. The NUS-48E corpus is composed of 48 sung and spoken (read) English songs, performed by 12 singers with different levels of musical training, where each singer speaks and sings four different songs from a selection of 20 unique songs. The data is humanly annotated at the phone-level, and has about 2800 annotated phones per subject, for each speech style.

First, we considered vowel duration differences (Figure 1a). It is seen that whereas spoken vowel durations are typically less than 300 milliseconds (ms), sung vowels are often over 500 ms and can extend over 1 second. Word recognition accuracy may be affected by the phone lengthening. In [19], it was found that the extended duration can lead to phone insertion and substitution errors, especially if the pitch is varying during the vowel. Further, vowel length is an important phonetic cue, (e.g., aiding discrimination between short and long vowel pairs such as /I/ vs /i:/). Artificially extending a vowel may therefore disrupt identification.

Figure 1b shows the vowel's fundamental frequency (F0) distribution for speech style and gender. Sung speech has a much larger F0 range than spoken speech. In contrast to spoken speech, which has virtually no overlap in F0 range between genders, the extended F0 ranges mean that the top end of the male range (200-300 Hz) overlaps with the lower female F0 values. For sung speech, different peaks can be seen in the male and female distributions. These peaks occur at the position of specific musical notes, e.g., some of the peaks in the male singer plot correspond to the notes E_3, $F_3^\#$ and $G_3^\#$. Notice that unlike spoken speech, which can be well represented with a single Gaussian in log frequency scale, the sung speech distribution would be better represented by a Gaussian mixture model where each component centred at a discrete note in the vocal range. At the highest sung F0 values vowel identification will become challenging as the sparseness of the harmonics means that the formant positions and bandwidths will be hard to determine.

We used the Kaldi 'probability of voicing' measure (POV) (Section 3), to examine the importance of the voicing information for the discrimination between voiced and unvoiced phones. Figure 2 shows the POV measure's probability distribution for the voiced/unvoiced fricative sounds (note, low POV implies more strongly voiced). For each condition, the Bhattacharyya distance (D_B) was measured to
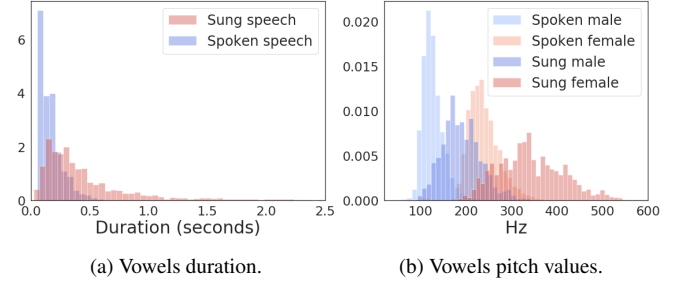

(a) Vowels duration.   (b) Vowels pitch values.

Fig. 1: Distribution of vowel duration (1a) and pitch value (1b), separated by sung and spoken speech styles.

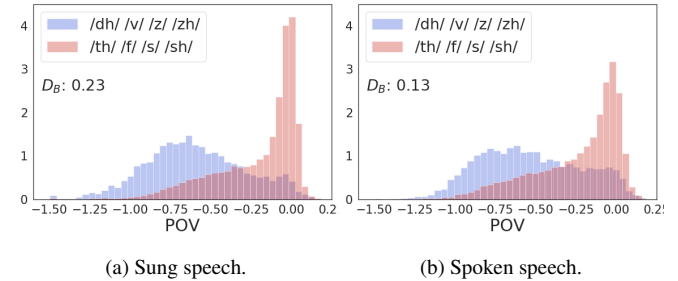(a) Sung speech.   (b) Spoken speech.

Fig. 2: Distribution of POV measure for voiced and unvoiced fricative phones. Bhattacharyya distance (D_B) is annotated in both plots.

show the similarity between the voiced and unvoiced distributions. For sung speech, although both classes show more voicing (i.e., a shift of the distributions to the left), the shift is greater for voiced phonemes leading to a higher D_B value. This suggests that the POV could be more informative for sung phonemes differentiation than for spoken speech. No variation on the D_B distance was observed for voiced versus unvoiced stop phonemes.

Finally, we analysed the voice quality (VQ) production differences between speech styles by using the fundamental frequency perturbance measurements jitter and shimmer, plus harmonic to noise ratio (HNR); parameters typically used for voice pathology detection [21]. These features have been proved useful for speaker recognition [22]. VQ measurements are divided into three subgroups; jitter, that measures the frequency variation from cycle-to-cycle; shimmer, that relates to the amplitude variation of the sound wave; and HNR, that measures the ratio between the harmonics and the glottal noise. Analysis using the NUS-48E showed that sung speech jitter and shimmer ranges are lower, and HNR is higher than in spoken speech. Among the factors that may explain these variations are vocal tract muscles "warm-up" during singing [23], interactions between jitter, shimmer and the degree of vibrato [24] and, the correlation between these parameters and the fundamental frequency, i.e., jitter will decrease at higher frequencies. The voice quality parameters ranges showed a clear gender distinction in spoken speech. In contrast, in singing, the gender distinction virtually disappears, only jitter maintains some degree of gender specification and, shimmer and HNR become more gender independent parameters. The weaker link to speaker identity may reduce the value of these parameters as speaker normalisation conditioning variables.

3. VOICE SOURCE FEATURES

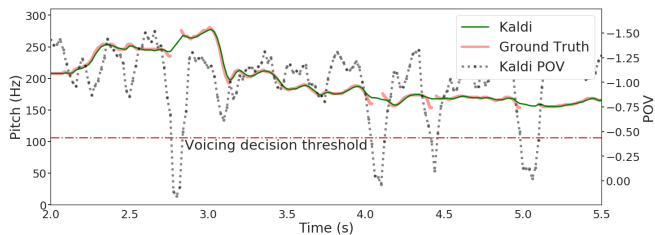This section describes the voice source-based features that will be evaluated in the ASR experiments.

6514

Fig. 3: Kaldi pitch tracker's pitch and POV estimation contrasted with the ground truth from 3.5 seconds from one MIR-1K excerpt.

3.1. Pitch and voicing estimation

For pitch feature extraction, we utilised the Kaldi pitch tracker [25], which is based on the getf0 (RAPT) algorithm [26]. Unlike getf0, Kaldi pitch tracker does not make a hard voicing decision; instead, it treats all frames as voiced sounds with an associated pitch, but with varying probability of voicing (POV). A Viterbi search is used to interpolate pitch estimates across unvoiced frames. The algorithm outputs a log pitch representation, and a normalised log pitch by using a short-time mean subtraction (as in [27]), a Gaussian-distributed POV feature and a two frame-context delta-pitch feature.

3.2. Tuning the Kaldi pitch tracker for sung speech

The Kaldi pitch tracker possesses a maximum pitch value (max-f0) and low-pass frequency cut-off (lowpass-cutoff) that are tuned for spoken speech. These were re-tuned for sung speech using the MIR-1K pitch annotated sung speech dataset (MIR-1K) [28]. MIR-1K is a collection of 1000 male and female song clips totalling 133 minutes, extracted from 110 karaoke songs selected from 5,000 Chinese pop songs. Pitch values have been manually annotated.

By employing a grid-searching technique, we evaluated seven values for max-f0 (between 400 and 1,000, with 100Hz step) and three values for lowpass-cutoff (1,000, 1,500 and 2,000). The performance was assessed by using the gross pitch error (GPE) [29] and fine pitch error (FPE) [30] measurements. GPE calculates the proportion of frames classified as voiced by both the estimation and ground truth, where the estimated pitch deviates more than one semi-tone from the ground truth [31]. However, as the Kaldi pitch tracker does not perform a hard voiced decision, we calculate the GPE using all the *voiced frames* from the ground truth. And, FPE is defined as the mean absolute error derived from the voiced frames where the reference deviates less than the GPE threshold [30].

Best performance was found with a max-f0 of 1,000Hz and lowpass-cutoff of 1,500Hz (compared to 400 and 1000 respectively for the original spoken speech tuning). Figure 3 shows an example of the predicted pitch by the Kaldi pitch tracker using these parameters, compared to the ground truth. Notice that Kaldi pitch tracker predicted the values with high accuracy and it interpolates between voiced frames to assign a pitch to the unvoiced ones. Additionally, this plot shows the POV estimation (right y-axis with inverted order), notice that in unvoiced areas, the POV value shows valleys to values below a voicing decision threshold estimated from MIR-1K.

3.3. Voice quality parameters

The VQ parameters were extracted using Praat [32]. First, we estimated jitter by analysing the timings of individual fundamental periods. In our experiments, we evaluated two jitter parameters, namely average absolute jitter (*jitta*) and relative average perturbation (*rap*) [21]. The *jitta* parameter measures the average *absolute* variation of

the fundamental period between consecutive cycles, and *rap* measures the *relative* difference between a period and the average of its two neighbours. Second, we estimated shimmer, defined as the average absolute difference between the *amplitude* of pairs of consecutive periods. Finally, HNR is a single parameter calculated from the auto-correlation function of the voice signal and is a measure of the degree of aperiodicity [21].

4. THE ASR BASELINE

4.1. The task

Recognition experiments have been performed using the DSing sung speech recognition corpus [8]. DSing is composed of the 4,460 karaoke performances, and is the English subset of the larger multi-language karaoke Smule Sing!300x30x2 dataset (Sing!) [33] released by Smule[2], in early 2018.

The performances are equally distributed by gender and organised by the country (30 in total) where the singer was located during the recording. Using the country information, DSing conveniently splits the data into three progressively larger training sets, named *DSing1* (15.1 hours), *DSing3* (44.7 hours) and *DSing30* (149.1 hours). The smallest training set DSing1 is constructed using 80% of the recordings from the UK. The remaining 20% is split into two balanced test sets, one for development (dev) and one for evaluation (eval) (1.5 hours in total). DSing3 extends DSing1 with recordings from Australia and the US. Although there is no information about the nationality of the singers, it can be safely assumed that most of the singers in DSing1 and DSing3 are native English speakers. The third and most extensive set, DSing30, covers all the English recordings from the 30 countries, including a large number of English recordings from non-native English speakers.

4.2. Baseline system description

As a baseline, we employ the sung speech recognition system from [8] built using the Kaldi ASR toolkit [34]. The acoustic features employed are 13 Mel frequency cepstral coefficients (MFCC) plus delta, delta-delta and energy, with 25 ms frame length and 15 ms of overlapping. Initial alignments are performed with a tri-phone speaker adapted Gaussian mixture model (GMM) and feature-space maximum likelihood linear regression (fMLLR). The tri-phone model is used to clean the training data using the standard Kaldi cleanup process[3]. A factorised time-delay neural network (TDNN-F) [35] acoustic model is then trained using 40 MFCCs with two frames context plus 100-dimensional i-Vectors, and a lattice-free maximum mutual information (LF-MMI) loss function [36].

A 3-gram MaxEnt language model (LM) is trained on an in-domain lyrics corpus, sourced from the Lyrics Wiki website[4]. This corpus is composed of the lyrics from songs from all artists in the DSing3 training set excluding songs that actually appear in DSing itself, plus the lyrics from all the artists from the Billboard The Hot 100 for the 31st December of the years 2015 to 2018. Language model rescoring is then performed with a 4-gram MaxEnt model trained on the same data. The vocabulary selected was the 26K most frequency words.

Using the 3-gram LM, the evaluation set performance was 42.28%, 28.67% and 22.32% for models trained with DSing1, DS-

---

[2]https://www.smule.com/

[3]https://github.com/kaldi-asr/kaldi/blob/master/egs/wsj/s5/steps/cleanup/clean_and_segment_data.sh

[4]https://lyrics.fandom.com/wiki/LyricWiki

6515

Table 1: Evaluation set performances (95% confidence interval WER) from experiments trained on DSing1, DSing3 and DSing30. Values in bold show WERs that are significantly better than the baseline.

Experiment	DSing1		DSing3		DSing30	
	3-gram	4-gram	3-gram	4-gram	3-gram	4-gram
Baseline	43.02 ± 0.55	38.14 ± 0.58	28.13 ± 0.14	24.40 ± 0.26	22.82 ± 0.21	19.88 ± 0.34
Kaldi LN	$\mathbf{40.99 \pm 0.49}$	$\mathbf{36.77 \pm 0.45}$	28.05 ± 0.24	24.27 ± 0.21	23.23 ± 0.28	19.87 ± 0.12
Kaldi L	$\mathbf{41.55 \pm 0.58}$	37.14 ± 0.52	27.80 ± 0.34	24.35 ± 0.27	22.92 ± 0.33	19.72 ± 0.25
Kaldi N	$\mathbf{41.61 \pm 0.37}$	37.40 ± 0.87	$\mathbf{27.79 \pm 0.27}$	24.28 ± 0.19	22.95 ± 0.15	19.67 ± 0.11
Kaldi LN + VQ	$\mathbf{41.17 \pm 0.30}$	$\mathbf{36.70 \pm 0.46}$	27.82 ± 0.26	$\mathbf{23.76 \pm 0.27}$	22.97 ± 0.32	19.60 ± 0.21

ing3 and DSing30 respectively, reducing to 37.63%, 24.27% and 19.60% WER after applying 4-gram LM rescoring. For full system details see [8].

5. EXPERIMENT AND RESULTS

5.1. Experiments

In order to establish the significance of our results we have repeated experiments multiple times. In particular, the AM training depends on the *random* initialisation of parameters, the *random* presentation order of training data, etc. Each training, though equally valid, can produce WER results that vary appreciably, and the variation can be mistaken for genuine performance variations. Therefore, for all experiments we retrain systems 11 times allowing us to treat the evaluation of the model performances statistically.

We first replicated the baseline system training it eleven times to calculate confidence intervals by using the mean and the standard error of the mean scores. After each training, the development set was used to select the LM-weight and words-insertion-penalty Kaldi's decoding parameters, and these parameters were then used to decode the evaluation set. These parameters were firstly estimated when decoding with the 3-gram LM and then when using the 4-gram LM for re-scoring.

Following the above procedure, we performed experiments by expanding the baseline MFCC + i-Vector feature vector using different voice source based feature combinations. The first experiment, **Kaldi LN**, evaluated the effect of including both Kaldi pitch representations; the log pitch and the normalised log pitch. The second, **Kaldi L**, evaluated the effect of using the log pitch, without the normalised log pitch. The third, **Kaldi N**, evaluated the effect of utilising only the normalised log pitch. For these three first experiments, both the delta pitch and POV were included. The final experiment expanded the best combination of MFCC plus pitch features – obtained from the results when training with the smallest DSing1 – with the four **VQ** features.

5.2. Results and analysis

Table 1 presents the system performances (WER) along with the 95% confidence intervals for the baseline system and the systems using the various combinations of voice source features. For each system, results are shown for each training dataset and for using each of the 3-gram or 4-gram LM. The baseline results for the 4-gram LM, 38.14 ± 0.58, 24.40 ± 0.26 and 19.88 ± 0.34 WER, for DSing1, DSing3 and DSing30 respectively, are consistent with the results reported in [8].

For the DSing1 model, using the four Kaldi pitch features (Kaldi LN experiment) reduces the error by about 2.0% ($p < .05$) 3-gram and, 1.4% ($p < .05$) 4-gram. A similar improvement is obtained by
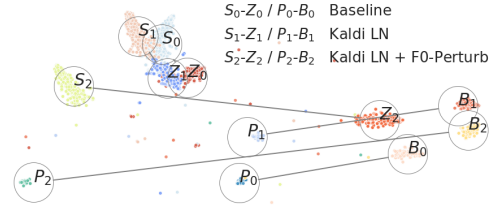

Fig. 4: T-SNE constructed with the posterior probabilities from model train on DSing30.

expanding Kaldi LN with the VQ features ($p < .05$). For the larger DSing3 training set, a reduction of 0.5% ($p = .057$) for 3-gram and, 0.7% ($p < .05$) 4-gram, is obtained when combining the pitch feature with the VQ. Voiced source features did not help to improve the models trained on the largest DSing30. The VQ features do not produce a significant improvement over using the pitch features, for any training set size.

Figure 4 shows a t-SNE plot of the \s\-\z\fricative and \p\-\b\plosive sounds, constructed by using the posterior probabilities from one model trained with DSing30. The subscript 0 represents the baseline model; subscript 1 Kaldi LN and, subscript 2 Kaldi LN + VQ. In the baseline model there are separate clusters for the voiced and unvoiced fricatives and plosive pairs, but the clusters overlap. When the model is informed with pitch features and VQ information, the distance between classes becomes wider and the classes are more compressed. Best discrimination is seen when the VQ features are included, giving some hint that they may be useful even though improvements in WER were not significant.

6. CONCLUSIONS

This paper presented an analysis of the differences in voice source characteristic between the sung and spoken speech styles, and its effect on a sung speech ASR performance. Using voice source features improved ASR performances, reducing WERs by between 0.7% to 1.4% on an existing state of the art baseline. The ASR results and t-SNE analysis suggested that voice source features help to compress the phoneme classes and to increase the distance between pairs of voiced/unvoiced phonemes. However, this effect only improves performances for ASR systems with the smaller training sets. When using the more extensive DSing30 (149.1 hrs) the vocal source based features were, surprisingly, found to provide no significant benefit for ASR performance. This suggests that, with enough training data sets, systems are able to learn the phonetic cues being carried in the voice source in a less direct manner (e.g., via the temporal dynamics of the MFCC features).

6516

7. REFERENCES

[1] A. Mesaros and T. Virtanen, "Recognition of phonemes and words in singing," in *Proc. IEEE ICASSP*, 2010.

[2] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2016.

[3] C. Tsai, Y. Tuan, and L. Lee, "Transcribing lyrics from commercial song audio: The first step towards singing content processing," in *Proc. IEEE ICASSP*, 2018.

[4] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[5] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE ICASSP*, 2015.

[6] K. J. Han, A. Chandrashekaran, J. Kim, and I. R. Lane, "The CAPIO 2017 conversational speech recognition system," *CoRR*, vol. abs/1801.00059, 2017.

[7] M. Rouvier and B. Favre, "Speaker adaptation of DNN-based ASR with i-vectors: Does it actually adapt models to speakers?" in *Proc. INTERSPEECH*, 2014.

[8] G. Roa Dabike and J. Barker, "Automatic lyric transcription from Karaoke vocal tracks: Resources and a baseline system," in *Proc. Interspeech*, 2019.

[9] H. Huang, Y. Hu, and H. Xu, "Mandarin tone modeling using recurrent neural networks," *arXiv:1711.01946*, 2017.

[10] M. M. Doss, T. A. Stephenson, and H. Bourlard, "Using pitch frequency information in speech recognition," in *Proc. EUROSPEECH*, 2003.

[11] G. Cloarec, D. Jouvet, and J. Monné, "Analysis of the modeling of pitch and voicing parameters for speaker-independent speech recognition systems," in *Speech Recognition and Intrinsic Variation (SRIV)*, 2006.

[12] M. Płonkowski and P. Urbanovich, "The use of pitch in large-vocabulary continuous speech recognition system," *Przeglad Elektrotechniczny*, vol. 92(8), pp. 78–81, 2016.

[13] S. Shahnawazuddin, A. Dey, and R. Sinha, "Pitch-adaptive front-end features for robust children's ASR," in *Proc. Interspeech*, 2016.

[14] S. Shahnawazuddin, R. Sinha, and G. Pradhan, "Pitch-normalized acoustic features for robust children's speech recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 8, pp. 1128–1132, 2017.

[15] J. Merrill and P. Larrouy-Maestri, "Vocal features of song and speech: Insights from Schoenberg's Pierrot lunaire," *Frontiers in Psychology*, vol. 8, no. JUL, 2017.

[16] A. D. Patel, M. Wong, J. Foxton, A. Lochy, and I. Peretz, "Speech intonation perception deficits in musical tone deafness (congenital amusia)," *Music Perception*, vol. 25, pp. 357–368, 04 2008.

[17] P. G. Vos and J. M. Troost, "Ascending and descending melodic intervals: Statistical findings and their perceptual relevance," *Music Perception: An Interdisciplinary Journal*, vol. 6, no. 4, pp. 383–396, 1989.

[18] J. Sundberg, *The science of the singing voice*. Northern Illinois University Press, 1987.

[19] D. Kawai, K. Yamamoto, and S. Nakagawa, "Lyric recognition in monophonic singing using pitch-dependent DNN," in *Proc. IEEE ICASSP*, 2017.

[20] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Proc. APSIPA ASC*, 2013.

[21] J. P. Teixeira, C. Oliveira, and C. Lopes, "Vocal acoustic analysis – jitter, shimmer and HNR parameters," *Procedia Technology*, vol. 9, pp. 1112–1122, 2013.

[22] M. Farrus, J. Hernando, and P. Ejarque, "Jitter and Shimmer Measurements for Speaker Recognition," in *Proc. INTERSPEECH*, 2007.

[23] C. Mezzedimi, M. Spinosi, T. Massaro, F. Ferretti, and J. Cambi, "Singing voice: acoustic parameters after vocal warm-up and cool-down," *Logopedics Phoniatrics Vocology*, pp. 1–9, 2018.

[24] J. Hakes, T. Shipp, and E. T. Doherty, "Acoustic characteristics of vocal oscillations: vibrato, exaggerated vibrato, trill, and trillo," *Journal of Voice*, vol. 1, no. 4, pp. 326–331, 1988.

[25] P. Ghahremani, B. Babaali, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for ASR," in *Proc. IEEE ICASSP*, 2014.

[26] D. Talkin and W. B. Kleijn, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[27] X. Lei, "Modeling lexical tones for Mandarin large vocabulary continuous speech recognition," Ph.D. dissertation, University of Washington, 2006.

[28] C. Hsu and J. R. Jang, "On the Improvement of Singing Voice Separation for Monaural Recordings Using the MIR-1K Dataset," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 18, no. 2, pp. 310–319, 2010.

[29] T. Drugman, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. INTERSPEECH*, 2011.

[30] M. Asgari and I. Shafran, "Improving the accuracy and the robustness of harmonic model for pitch estimation," in *Proc. INTERSPEECH*, 2013.

[31] O. Babacan, T. Drugman, N. d'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *Proc. IEEE ICASSP*, 2013.

[32] P. Boersma, "A system for doing phonetics by computer," *Glot International*, vol. 5:9/10, pp. 341–345, 11 2000.

[33] Smule Sing!300x30x2 Dataset, "https://ccrma.stanford.edu/damp/," accessed September 2018.

[34] D. Povey *et al.*, "The Kaldi speech recognition toolkit," in *IEEE ASRU*, 2011.

[35] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. Interspeech*, 2018.

[36] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016.