# The NUS Sung and Spoken Lyrics Corpus: A Quantitative Comparison of Singing and Speech

Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim and Ye Wang
School of Computing, National University of Singapore, Singapore.
E-mail: [zhiyan, fanght, li-bo, simkc, wangye]@comp.nus.edu.sg

***Abstract*— Despite a long-standing effort to characterize various aspects of the singing voice and their relations to speech, the lack of a suitable and publicly available dataset has precluded any systematic study on the quantitative difference between singing and speech at the phone level. We hereby present the NUS Sung and Spoken Lyrics Corpus (NUS-48E corpus) as the first step toward a large, phonetically annotated corpus for singing voice research. The corpus is a 169-min collection of audio recordings of the sung and spoken lyrics of 48 (20 unique) English songs by 12 subjects and a complete set of transcriptions and duration annotations at the phone level for all recordings of sung lyrics, comprising 25,474 phone instances. Using the NUS-48E corpus, we conducted a preliminary, quantitative study on the comparison between singing voice and speech. The study includes duration analyses of the sung and spoken lyrics, with a primary focus on the behavior of consonants, and experiments aiming to gauge how acoustic representations of spoken and sung phonemes differ, as well as how duration and pitch variations may affect the Mel Frequency Cepstral Coefficients (MFCC) features.**

## I. INTRODUCTION

In audio signal analysis, it is important to understand the characteristics of singing voice and their relation to speech. A wide range of research problems, such as singing and speech discrimination and singing voice recognition, evaluation, and synthesis, stand to benefit from a more precise characterization of the singing voice. Better solutions to these research problems could then lead to technological advancements in numerous application scenarios, from music information retrieval and music edutainment to language learning [1] and speech therapy [2].

Given the similarity between singing and speech, many researchers classify the former as a type of the latter and try to utilize the well-established automatic speech recognition (ASR) framework to handle the singing voice [3][4][5]. Due to the lack of phonetically annotated singing datasets, models have been trained on speech corpora and then adapted to the singing voice. This approach, however, is intrinsically limited because the differences between singing and speech signals are not captured. Thus, a quantitative comparison of singing voice and speech could potentially improve the capability and robustness of current ASR systems in handling singing voice.

Despite long-standing efforts to characterize various aspects of singing voice and their relations to speech [7][8], the lack of a suitable dataset has precluded any systematic quantitative study. Given that most existing models are statistics-based, an ideal dataset should not only have a large size but also exhibit sufficient diversity within the data. To explore the quantitative differences of singing and speech at the phone level, the research community needs a corpus of phonetically annotated recordings of sung and spoken lyrics by a diverse group of subjects.

In this paper, we introduce the NUS Sung and Spoken Lyrics Corpus (NUS-48E corpus for short), 48 English songs the lyrics of which are sung and read out by 12 subjects representing a variety of voice types and accents. There are 20 unique songs, each of which covered by at least one male and one female subject. The total length of audio recordings is 115 min for the singing data and 54 min for the speech data. All singing recordings have been phonetically transcribed with duration boundaries, and the total number of annotated phones is 25,474. The corpus marks the first step toward a large, phonetically annotated dataset for singing voice research.

Using the new corpus, we conducted a preliminary study on the quantitative comparison between sung and spoken lyrics. Unlike in speech, the durations of syllables and phonemes in singing are constrained by the music score. They have much larger variations and often undergo stretching. While vowel stretching is largely dependent on the tempo, rhythm, and articulation specified in the score, consonant stretching is much less well understood. We thus conducted duration analyses of the singing and speech data, primarily focusing on consonants. We also hope to better understand how one can borrow from and improve upon the state-of-the-art speech processing techniques to handle singing data. We thus carried out experiments to quantify how the acoustic representations of spoken and sung phonemes differ, as well as how variations in duration and pitch may affect the Mel Frequency Cepstral Coefficients (MFCC) features. The results of both the duration and spectral analyses are hereby presented and discussed.

The remainder of this paper is organized as follows. Section II provides an overview of existing datasets and related works on the differences between singing and speech. Section III describes the collection, annotation, and composition of the NUS-48E corpus. Section IV and V present the results of the duration analyses and spectral comparisons, respectively. Finally, Section VI and VII conclude this paper and suggest future work.

## II. RELATED WORK

### A. Singing Voice Dataset

Singing datasets of various sizes and annotated contents are available for research purposes. To the best of our knowledge, however, none has duration annotations at the phoneme level.

Mesaros and Virtanen conducted automatic recognition of sung lyrics using 49 singing clips, 19 of which are from male singers and 30 from female singers [4]. Each clip is 20-30 seconds long, and the complete dataset amounts to approximately 30 minutes. Although a total of 4770 phoneme instances are present, the lyrics of each singing clip are manually transcribed only at the word level, without any duration boundaries.

The MIR-1K dataset [6] is a larger dataset consisting of 1000 clips from 110 unique Chinese songs as sung by 19 amateur singers, 8 of whom female. The total length of the singing clips is 133 minutes. Since this dataset is intended for singing voice separation, annotations consist of pitch, lyrics, unvoiced frame types, and vocal/non-vocal segmentation, but do not contain segmentation on the word level or below.

AIST Humming Database (AIST-HDB) [9] is a large database for singing and humming research. The database contains a total of 125.9 hours of humming/singing/reading materials, recorded from 100 subjects. Each subject produced 100 excerpts of 50 songs chosen from the RWC Music Database (RWC-MDB) [16]. While the lyrics of the songs are known, neither the AIST-HDB nor the RWC-MDB provides any phoneme or word boundary annotation.

### B. Differences of Singing and Speech

Observations on differences of singing and speech have been reported and studied [7][8]. The three main differences lie in phoneme duration, pitch, and power. Constrained by the music score and performance conventions, the singing voice stretches phonemes, stabilizes pitches, and roams within a wider pitch range. The power changes with pitch in singing but not in speech.

Ohishi et al. studied the human capability in discriminating singing and speaking voices [10]. They reported that human subjects could distinguish singing and speaking with 70.0% accuracy for 200-ms signals and 99.7% for one-second signals. The results suggest that both temporal characteristics and short-term spectral features contribute to perceptual judgment. The same research group also investigated short-term MFCC features and long-term contour of the fundamental frequency (F0) in order to improve machine perform on singing-speaking discrimination [8]. F0 contour works better for signals longer than one second, while MFCC performs better on shorter signals. The combination of the short-term and long-term features achieved more than 90% accuracy for two-second signals.

Since singing and speech are similar from various aspects, finding the right set of features to discriminate the two is crucial. A set of features derived from harmonic coefficient and its 4Hz modulation values are proposed in [11]. While in [12], a feature selection solution among 276 features is introduced.

### C. Conversion between Singing and Speech

The conversion between speaking and singing has also attracted research interest. A system for speech-to-singing synthesis is described in [13]. By modifying the F0, phoneme duration, and spectral characteristics, the system can synthesize a singing voice with naturalness almost comparable to a real singing voice using a speaking voice and the corresponding text as input. A similar system is developed in [14] to convert spoken vowels into singing vowels. On the other hand, the SpeakBySinging [15] system converts a singing voice into a speaking voice while retaining the timbre of the singing voice.

## III. The NUS Sung and Spoken Lyrics Corpus

### A. Audio Data Collection

**Song Selection.** We selected twenty songs in English as the basis of our corpus (see Table I). They include well-known traditional songs and popular songs that have been regional and international hits, as well as several songs that may be less familiar but are chosen for their phonemic richness and ease of learning. In addition, lyrics of some songs, such as *Jingle Bells* and *Twinkle Twinkle Little Star*, are expanded to include verses other than the most familiar ones to further enhance the phonemic richness of the corpus, while overly repetitive lines or instances of scat singing, such as those found in *Far Away from Home* and *Lemon Tree*, are excised to better preserve phonemic balance. The list of songs and their selected lyrics are posted on our study website[1].

**Subject Profile.** We recruited 21 subjects, 9 males and 12 females, from the National University of Singapore (NUS) Choir and the amateur vocal community at NUS. All subjects are enrolled students or staff of the university. They are 21 to 27 years of age and come with a wide range of musical exposure, from no formal musical training to more than 10 years of vocal ensemble experience and vocal training. All four major voice types (soprano, alto, tenor, and bass) are represented, as well as a spectrum of English accents, from North American to the various accents commonly found in Singapore. Local accents tend to be less apparent in singing than in speaking, a phenomenon that becomes more marked as the subject's vocal experience increases. Subjects are all proficient speakers of English, if not native speakers.

**Collection Procedure.** Subjects visited the study website to familiarize with the lyrics of all twenty songs before coming to our sound-proof recording studio (STC 50+) for data collection. An Audio-Technica 4050 microphone with pop filter was used for the recording. Audio data were collected at 16-bit and 44.1kHz using Pro Tools 9, which also generated a metronome with downbeat accent to set the tempi and to serve as a guide for singing. The metronome was fed to the subject via the headphone. The selected lyrics for all songs were printed and placed on a music stand by the

[1] http://singingevaluation.wordpress.com/2012/11/22/songs-to-pick/

TABLE I
SONGS IN THE NUS CORPUS

| Song Name | Artist / Origin (Year) | Tempo (bpm) | Audio Length Estimate (s) | Phone Count Estimate | Female Subject[a] | Male Subject[a] |
|---|---|---|---|---|---|---|
| Edelweiss | The Sound of Music (1959) | Med (96) | 65 | 140 | 03 | 11 |
| Do Re Mi | The Sound of Music (1959) | Fast (120) | 67 | 280 | 05 | 08 |
| Jingle Bells | Popular Christmas Carol | Fast (120) | 85 | 630 | 05 | 08 |
| Silent Night | Popular Christmas Carol | Slow (80) | 165 | 340 | 01 | 09 |
| Wonderful Tonight | Eric Clapton (1976) | Slow (80) | 180 | 450 | 02 & 06 | 07 & 10 |
| Moon River | Breakfast at Tiffany's (1961) | Slow (88) | 160 | 380 | 05 | 08 |
| Rhythm of the Rain | The Cascades (1962) | Med (116) | 85 | 460 | 04 | 12 |
| I Have a Dream | ABBA (1979) | Med (112) | 135 | 390 | 06 | 10 |
| Love Me Tender | Elvis Presley (1956) | Slow (72) | 140 | 310 | 03 | 11 |
| Twinkle Twinkle Little Star | Popular Children's Song | Fast (150) | 115 | 640 | 01 | 09 |
| You Are My Sunshine | Jimmy Davis (1940) | Slow (84) | 167 | 620 | 02 & 06 | 07 & 10 |
| A Little Love | Joey Yung (2004) | Slow (84) | 53 | 250 | 01 | 09 |
| Proud of You | Joey Yung (2003) | Slow (84) | 140 | 680 | 03 | 11 |
| Lemon Tree | Fool's Garden (1995) | Fast (150) | 160 | 900 | 05 | 08 |
| Can You Feel the Love Tonight | Elton John (1994) | Slow (68) | 175 | 540 | 04 & 06 | 10 & 12 |
| Far Away from Home | Groove Coverage (2002) | Med (112) | 140 | 680 | 04 | 12 |
| Seasons in the Sun | Terry Jacks (1974); Westlife (1999) | Med (100) | 175 | 920 | 01 | 09 |
| The Show | Lenka (2008) | Fast (132) | 200 | 980 | 03 | 11 |
| The Rose | Bette Midler (1979) | Slow (68) | 175 | 450 | 02 | 07 |
| Right Here Waiting | Richard Marx (1989) | Slow (88) | 160 | 550 | 02 & 04 | 07 & 12 |

[a] Number in these columns are code identifications of subject singers. See Table III

microphone for the subject's reference. Except metronome beats heard through the headphone, no other accompaniment was provided, and subjects were recorded a cappella.

For each song, the selected lyrics were sung first. While the tempo was set, the subject could choose a comfortable key and were free to make small alterations to rhythm and pitch. Then, the subject's reading of the lyrics was recorded on a separate track. When a track with all the lyrics clearly sung or spoken was obtained, the subject proceeded to the next song. A few pronunciation errors were allowed as long as the utterance remained clear. Except the occasional rustles of the lyric printouts, miscellaneous noise was avoided or excised from the recording as much as possible.

For each subject, an average of 65 minutes of audio data was thus collected in 20 singing (~45min) and 20 reading tracks (~20min). Each track was then bounced from Pro Tools as a wav file for subsequent storage, annotation, and audio analyses. At the end of the recording session, we reimbursed each subject with a S$50 gift voucher for the university co-op store.

### *B. Data Annotation*

We adopted the 39-phoneme set used by the CMU Dictionary (see Table II) for phonetic annotation [17]. Three annotators used Audacity to create a label track for each audio file, and labeled phones and their timestamps are exported as a text file. Phones were labeled not according to their dictionary pronunciation in American English but as they had been uttered. This was done to better capture the effect of singing as well as the singer's accent on the standard pronunciation. We also included two extra labels, *sil* and *sp*, to mark the lengths of silence or inhalation between words (and, occasionally, between phones mid-word) and all duration-less word boundaries, respectively (see Fig. 1). Labels of one annotator were checked by another to ensure inter-rater consistency.

### *C. Corpus Composition*

Due to the time-consuming nature of phonetic transcription and the limitations on manpower, for the first version of the corpus we only manually annotated the singing data of 12 subjects. They include 6 males and 6 females and represent all voice types and accent types (see Table III). For each subject, we selected 4 songs to annotate. To ensure that all 20 songs were annotated at least once for both genders and that the number of annotated phones for each subject remained roughly equal, we ranked the songs by the number of phones estimated using the CMU Dictionary and assigned them accordingly (see Table I). At this stage, each subject has

TABLE II
PHONEMES AND PHONEME CATEGORIES

| Class | CMU Phonemes |
|---|---|
| Vowels | AA, AE, AH, AO, AW, AY, EH, ER, EY, IH, IY, OW, OY, UH, UW |
| Semivowels | W, Y |
| Stops | B, D, G, K, P, T |
| Affricates | CH, JH |
| Fricatives | DH, F, S, SH, TH, V, Z, ZH |
| Aspirates | HH |
| Liquids | L, R |
| Nasals | M, N, NG |

around 2100 phones annotated, and the corpus contains a total of 25,474 phone instances.

Annotation for spoken lyrics is generated by aligning the manually-labeled phone strings of the sung lyrics to the spoken lyrics using conventional Gaussian Mixture Model (GMM) – Hidden Markov Model (HMM) system trained on the Wall Street Journal (WSJ0) corpus (see Sec. 5 for details). While numerous discrepancies might exist between the actual sung and spoken versions, arising from the articulatory peculiarities of subjects and the differing methods of alignment, the annotated spoken lyrics allow us to make broad and preliminary observations about the extent of phonemic stretching between sung and spoken lyrics. As part of our future work, we will expand our corpus to include manual annotations of the spoken lyrics.

## IV. DURATION ANALYSIS

### A. Consonants Stretching

In singing, vowels are stretched to maintain musical notes for certain durations, and their durations are to a large extent dictated by the score. While the stretching of vowels is much more pronounced, consonants are nevertheless stretched at a non-trivial level (see Fig. 2). As the factors influencing consonant duration are less apparent than those for vowels, we will explore not only how much stretching takes place but also what may affect the amount of stretching.

TABLE III
SUBJECTS IN THE NUS CORPUS

| Code | Gender | Voice Type | Sung Accent | Spoken Accent |
|---|---|---|---|---|
| 01 | F | Soprano | North American | North American |
| 02 | F | Soprano | North American | North American |
| 03 | F | Soprano | North American | Mild Local Singaporean |
| 04 | F | Alto | Mild Malay | Mild Malay |
| 05 | F | Alto | Malay | Malay |
| 06 | F | Alto | Mild Malay | Mild Malay |
| 07 | M | Tenor | Mild Local Singaporean | Mild Local Singaporean |
| 08 | M | Tenor | Northern Chinese | Northern Chinese |
| 09 | M | Baritone | North American | North American |
| 10 | M | Baritone | North American | Standard Singaporean |
| 11 | M | Baritone | North American | North American |
| 12 | M | Bass | Local Singaporean | Local Singaporean |

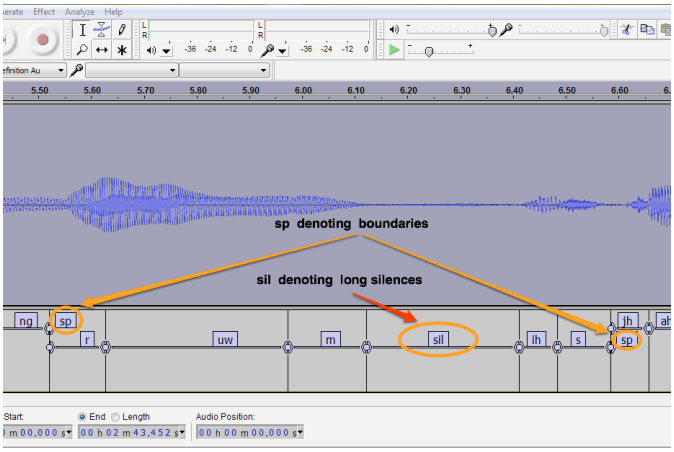

Fig. 1 SIL and SP labels denoting boundaries

The stretching ratio is computed as follows,

$$sr = T_{singing} / T_{speech} , \quad (1)$$

where $sr$ is the stretching ratio and $T_{singing}$ and $T_{speech}$ the durations of the phoneme in singing and the corresponding speech, respectively. The higher the $sr$ value, the more the phoneme is stretched in singing.

In the speech-to-singing conversion system developed in [13], the authors use fixed ratios for different types of consonants. The ratios used are experimentally determined from observations of singing and speech signals. Using the NUS-48E corpus, we analyzed the stretching ratio of

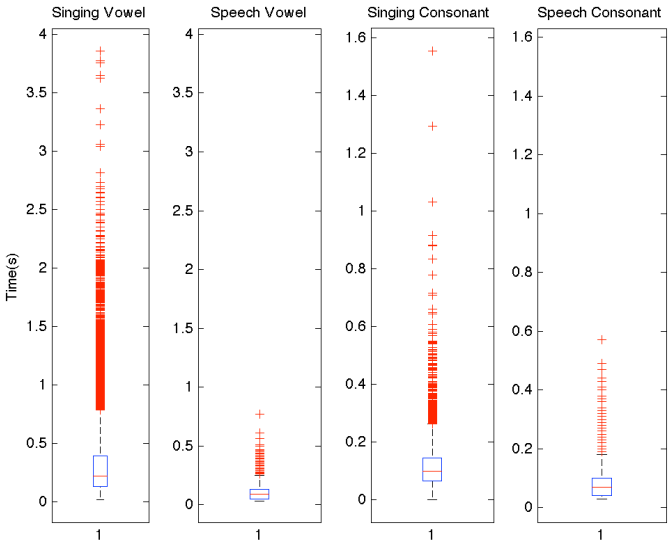

Fig. 2 Comparison on Duration Stretching of Vowels and Consonants in singing

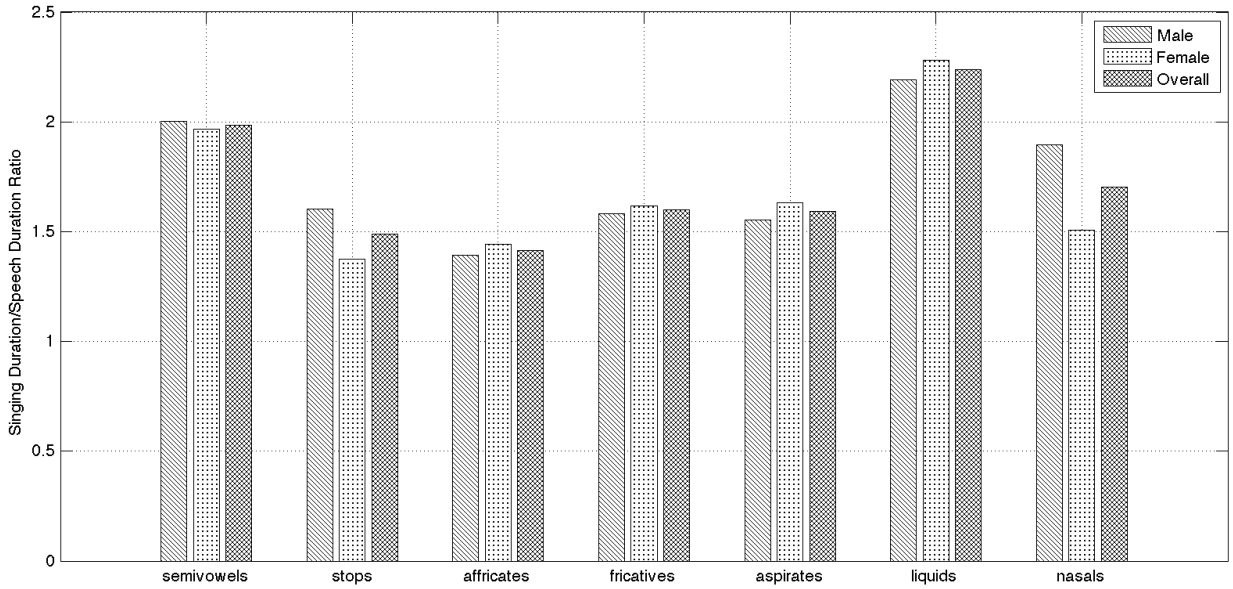


Fig. 3 Average Stretching Ratios of Seven Types of Consonants

consonants. Given that the phoneme alignment on speech data is automatically generated with a speech recognizer while the phoneme boundaries on singing data are manually annotated, the results remain preliminary observations.

As shown in Fig. 3, among the 7 types of consonants compared, liquids, semivowels, and nasals exhibit larger stretching ratios (2.2371, 1.9852, and 1.7027, respectively.) This result conforms to the intuition that these types of sonorants could be sustained and articulated for a longer period of time than others types such as stops and affricates.

Another interesting question is how the consonants are stretched in syllables of various lengths. The length of the syllables may have an effect on the length of consonants. As shown in Fig. 4, when syllable length starts to grow, the stretching ratio of semivowels increases accordingly. After the syllable length reaches around 1 second, however, the stretching ratio of semivowels tends to decrease. Not surprisingly, since vowels are the dominant constituent of syllables, the stretching ratio of vowels keeps growing when syllables become longer.

Observations on other types of consonants are similar to that discussed above for semivowels.

### *B. Subject Variations on Consonants Stretching*

As observations in the previous section only describe an overarching trend across all consonants for all subjects, it is important to check whether individual subjects follow such a trend consistently. We first investigated the differences with respect to gender. Fig. 5 shows the probability density functions (PDF) for the stretching ratios of both gender groups. The difference between them is negligible, suggesting that consonant stretching ratio is gender-independent. Next, we compared individual subjects. For example, subjects 05 and 08 contributed the same 4 songs, *Do Re Mi*, *Jingle Bells*, *Moon River*, and *Lemon Tree*. Subject 05 is a female with Malay accent and has had two years of choral experience at the time of recording, while subject 06 is a male with northern Chinese accent and had no vocal training whatsoever. As Fig. 6 shows, the distributions of the consonant stretching ratios of the two subjects remain roughly the same despite individual differences in accent and musical exposure. Therefore, the extent of consonant stretching may be attributed more to the act of singing itself than any discrepancy in the vocal practice of the singers.

### *C. Syllabic Proportions of Consonants*

Syllabic proportions are calculated as quotients of the consonant durations and the syllable durations. A phone with longer duration might not take up a higher proportion as it may be part of a long syllable.

Figure 7 shows the syllabic proportion for all consonant types and both gender groups. Overall, semivowels have the highest proportion while aspirates and stops have the lowest. With aspirates as the lone exception, the syllabic proportions of all consonant types are higher in males than in females.

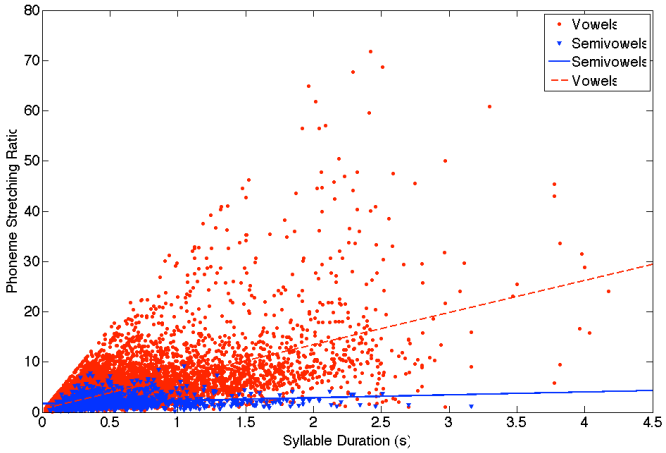


Fig. 4 Comparison on Duration Stretching Ratio across Different Length of Syllables for Vowels and Semivowels

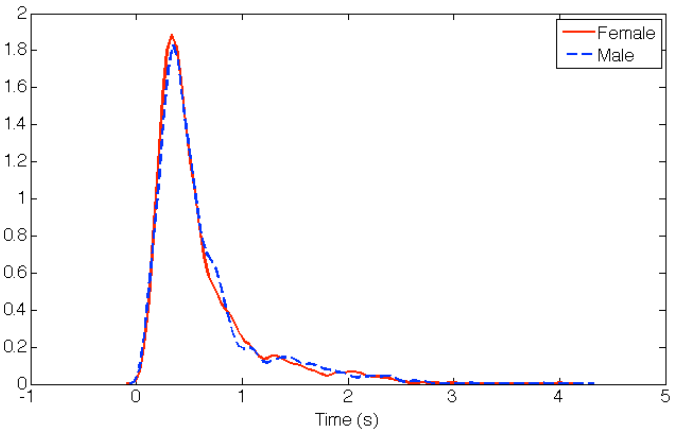


Fig. 5 Comparison on Probability Density Function of Consonants Duration Stretching Ratio with Respect to Gender

Further observation confirms that the absolute duration lengths of both consonants and syllables are larger in male subjects. This is an unexpected and interesting phenomenon given the observation made in the last subsection, namely that consonant durations seem to be stretched to similar extents in subjects of both genders.

Three factors could contribute to such a phenomenon. First, male and female subjects may have somewhat different duration distributions for consonants and vowels within the spoken syllables to begin with. Second, the stretching of sung vowels could exhibit gender-related discrepancies. Lastly, structure of the same syllable in the lyrics could be different between speech and singing, especially for subjects who would sing in a different accent. A dropped consonant or a diphthongized vowel could alter syllable makeup and affect syllable length. Once we have expanded our corpus to include phonetic annotations for the spoken lyrics, we plan to further our comparison study to examine these factors.

### *D. Consonant Position and its Effect on Proportion*

Within a syllable, consonants may appear at different positions. For example, in the word *love* (/l/ /ah/ /v/), consonant /l/ is located at the beginning of the word; while in *soul* (/s/ /ow/ /l/), it is at the end. We are interested to see whether this positioning has any effect on the syllabic proportion. We first defined four consonant positions:

1. **Starting**: at the beginning of a word, e.g. /g/ in *go*
2. **Preceding**: preceding a vowel, but not at the beginning of a word, e.g. /m/ in *small*

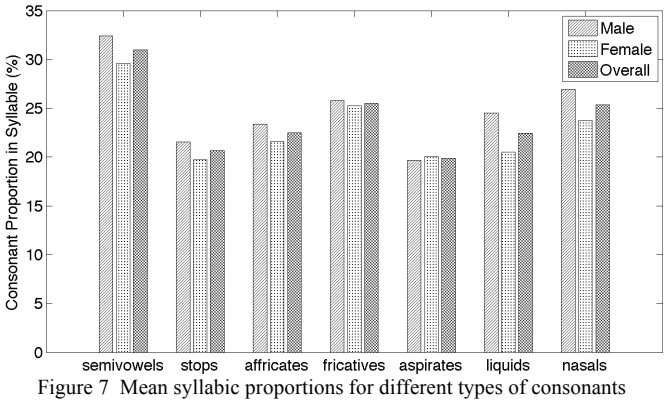


Figure 7 Mean syllabic proportions for different types of consonants

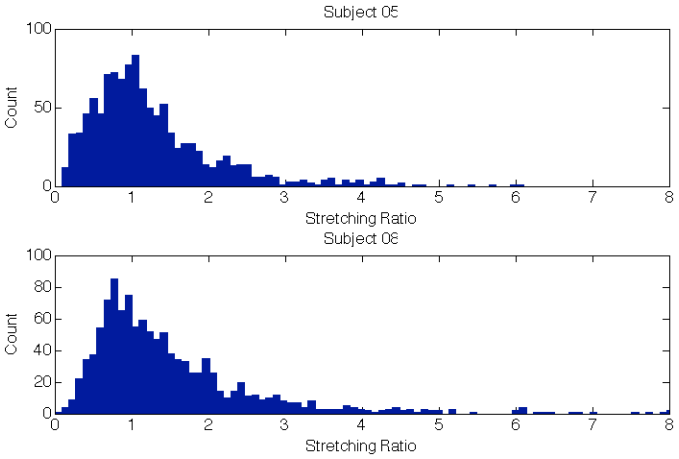


Fig. 6 Comparison on Consonants Duration Stretching Ratio of Subject 05 and Subject 08

3. **Succeeding**: succeeding a vowel, but not at the end of a word, e.g. /l/ in *angels*
4. **Ending**: at the end of a word, e.g. /t/ in *at*

We compared the syllabic proportions for the seven consonant categories with respect to positioning. The results are shown in Fig. 8. Semivowels and stops at the starting position are much more prominent than those at the end, while the opposite is observed for affricates and nasals. The syllabic proportions of fricatives, aspirates and liquids are largely similar between the starting and ending position.

For all consonants, the proportion for preceding position is significantly lower than that of the starting one. The phenomenon is mirrored for the succeeding and ending positions, in which the latter is much more prominent than the former.

## V. Spectral Analysis

Although we could build a conventional Gaussian Mixture Model (GMM) – Hidden Markov Model (HMM) system using the NUS-48E corpus, the performance is expected to be low mainly due to the following two factors: the limited amount of speech data and the variation of accents among the subjects. While few large, high-quality singing corpora are available for academic research, there are numerous standard speech corpora. We adopted the Wall Street Journal (WSJ0) corpus, a large collection of read speech with texts drawn from a machine-readable corpus of Wall Street Journal news,

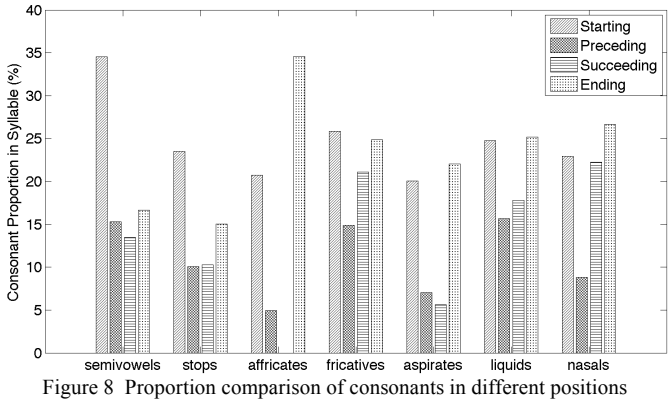


Figure 8 Proportion comparison of consonants in different positions

to train our speech GMM-HMM system, which is built to maximize the likelihood of the training data using the Hidden Markov Model Toolkit (HTK). The system adopts the CMU phoneme set used in the NUS-48E corpus and has a total of 2419 tied triphone states to model the various phoneme realizations in different acoustic contexts. Each state is modeled by a GMM with 16 components. On the benchmark 5k close vocabulary speech recognition task, our model has a word error rate (WER) of 5.51% when decoding with the bigram language model.

For comparison purpose, we also built a simple monophone based GMM-HMM system using the singing data of the NUS-48E corpus. Instead of the doing automatic alignment of the training data, we fixed the phoneme boundary according to the human annotations during training. Similarly, this singing GMM-HMM system also has 16 Gaussian for each state.

Both the speaking and singing waveform signals are processed with a 25ms time window and a shift of 10ms. Twelve dimensional MFCC features together with an energy term are extracted from each time window. These 13 terms, along with their first order and second order derivatives, make up the final, 39-dimensional feature vector.

### A. *Phoneme likelihood score comparison*

The GMM-HMM system trained on the WSJ0 corpus captures the spectral characteristics of the speech signals, and we used it to perform alignment on both the speech and singing data in the NUS-48E corpus. The alignment on singing data was restricted with the manually labeled boundaries. During both alignment tasks, the likelihood score generated by the GMM-HMM system were stored. Since the system is trained on a speech corpus, it is expected to perform worse on singing data. However, the difference between the likelihood scores of singing and speech phonemes carries useful information. It can serve as an indirect measure of the distance between the acoustic representation of the singing phoneme and that of the speech phoneme, i.e. a higher difference between the likelihood scores implies greater discrepancy between the acoustic characteristics of the two signals.

The likelihood score for each phoneme is a cumulative score on all frames contained in that phoneme. As durations of different phones vary significantly, the cumulative scores could be misleading. Thus we use the average likelihood score, which is computed by dividing the cumulative score by the frame count.

Then, we define the *Likelihood Difference (LD)* as

$$LD = abs(ALS_{singing} - ALS_{speech}), \quad (2)$$

where $ALS_{singing}$ and $ALS_{speech}$ are the average likelihood score for the singing phoneme and speech phoneme, respectively. As we only wished to gauge the extent of the likelihood differences, the absolute value of the difference is used to avoid negative scores cancelling out positive ones.

The comparison of likelihood differences between singing and speech phonemes of all phoneme types are shown in Fig. 9. Results show that females have higher likelihood differences for all phoneme types, especially liquids, which implies that there may be more differences in terms of accoustic features on female singing.

The likelihood differences of affricates and fricatives are lower than the other categories, suggesting that the accoustic features of these two phoneme types may be more similar between singing and speech.

While the 39-dimentional MFCC feature vector preserves the identity of the phoneme in question, it might have neglected information indicative of the difference between singing and speech. Therefore, likelihood difference is by no means a definitive measure on the differences of singing and speech phonemes. However, our observations may provide clues for further studies.

### B. *Understanding the effects of duration on MFCCs*

As variations in phoneme duration is one of the major differences between speaking and singing, we conducted preliminary experiments to see if they affect the MFCC features commonly used for speech analysis.

For simplicity, we converted duration into a discrete variable by dividing its whole value range into 10 bins with equal cumulative probability mass, i.e. each bin contains around 10% of the samples. Binning is carried out for each and every phoneme. We then estimate a single Gaussian to model the MFCC feature distribution for each bin of the phoneme. Ideally, there should be 390 different models, i.e. 39 phonemes each having 10 duration bins. Because the sung and spoken instances of a phoneme are binned together, the duration range of the sung instances could make it so that the spoken instances might not be distributed into all 10 bins, and vice versa. In the end, we obtained 348 separate models for speech and 366 for singing.

We then built decision trees to cluster these models together by asking questions based on the durations. For each phoneme, the 10 bins require 9 boundary values to split and hence 9 questions on the decision tree. The speech models and singing models are clustered separately. Clustering is carried out at each step by selecting the question that increases the data likelihood the most. If changes in a phoneme's MFCC features are affected by its duration, it would be more difficult to reduce the number of model clusters across the duration range, resulting in a lower reduction rate after clustering. After the decision tree

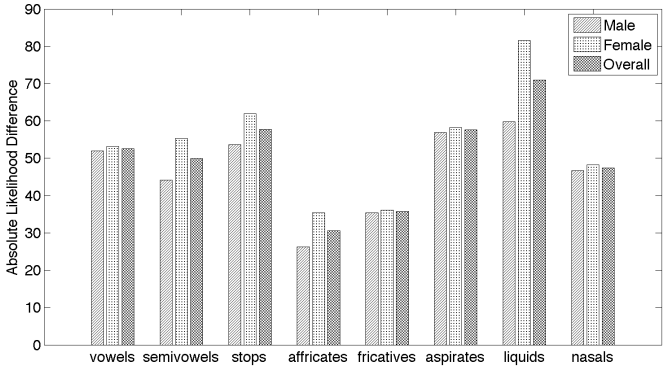

Fig. 9 Mean Differences of Likelihood Scores for All Phoneme Categories

clustering, we obtained 140 clusters for speech models and 177 clusters for singing models. The relative model reduction rate is 59.78% and 51.64%, respectively.

### C. *Understanding the effects of pitch on MFCCs*

We conducted the same set of experiments to evaluate the effects of pitch on the MFCC features. We also used 10 bins to discretize the pitch values and to ensure that all the bins have balanced cumulative density masses. After binning, we obtained 334 models for speech and 342 for singing. After decision tree building and clustering, the number of models was reduced to 182 for speech and 259 for singing, yielding reduction rates of 45.51% and 24.27%, respectively. The reduction rate for singing data is much lower than that of speaking data, especially when compared to the duration based clustering, suggesting that pitch differences can bring more variations to MFCC features.

## VI. Conclusion

In this paper, we introduce the NUS Sung and Spoken Lyrics Corpus (NUS-48E Corpus), which is an ongoing effort toward a comprehensive, well-annotated dataset for singing voice related research. The corpus contains: 12 subjects representing various accents and extents of musical background; 48 songs with reasonably balanced phoneme distribution. To the best of our knowledge, the NUS-48E corpus is the first singing voice dataset to offer annotations on the phone level.

Using our corpus, we conducted a comparative study of sung and spoken lyrics. Specifically, we investigated the duration and spectral characteristics of the phonemes in singing and speech. A preliminary analysis on the stretching ratio of sung phonemes is presented. Differences among stretching ratios of seven consonant categories are compared and the variations among subjects discussed. We investigated the syllabic proportion of consonants in sung lyrics with respect to consonants types as well as consonant positions within the syllable. Using a GMM-HMM system trained on a large speech corpus, we studied the difference between singing and speech phonemes in terms of MFCC features. The effects of duration and pitch on acoustic features are also discussed. The level of difference was measured through *Likelihood Difference*, which is based on the likelihood score generated by the GMM-HMM system. The effects of duration and pitch on MFCC features are examined by clustering acoustic models with decision trees.

## VII. Future Work

While the NUS-48E corpus contains only 48 annotated songs due to limitations on time and qualified manpower, we have recorded a total of 420 song samples (21 subjects, each singing all 20 songs in Table I). On the one hand, we will continue to enlarge our corpus by annotating the remaining songs. On the other hand, we will begin annotating the spoken data in order to provide the ground truth for future comparison studies. Using the enlarged corpus, we would also like to repeat some of the works mentioned in Section II to provide quantitative verifications for the observations reported in the literature.

As the comparison study presented in this paper is preliminary in nature, its results could be further explored and analyzed. Subsequent experiments will aim to answer the question and test the theory raised by the current observations, such as the differing syllabic proportions of consonants in male subjects. In the process, we hope to unearth new observations and raise new questions that could advance the community's understanding of the relationship between singing voice and speech. Eventually, we seek to combine the knowledge gained from the corpus and the literature to better adapt state-of-the-art speech evaluation technologies for the singing voice.

## Acknowledgment

The authors deeply appreciate the assistance of Kenny Yang and Amelia Dizon with phonetic annotation.

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

## References

[1] S. L. Medina, "The Effects of Music upon Second Language Vocabulary Acquisition," *Annual Meeting of the Teachers of English to Speakers of Other Languages*, March 1990.

[2] M. L. Albert, R. W. Sparks, and N. A. Helm, "Melodic intonation therapy for aphasia," *Arch. Neurol.*, vol. 29, issue 2, pp. 130-131, August 1973.

[3] M. Mehrabani and J. H. Hansen, "Singing speaker clustering based on subspace learning in the GMM mean supervector space," *Speech Commun.*, vol. 55, issue 5, pp. 653-666, June 2013.

[4] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP J. Audio Speech Music Process.*, vol. 2010, February 2010.

[5] A. Loscos, P. Cano, and J. Bonada, "Low-delay singing voice alignment to text," *Proc. Int. Comput. Music Conf.*, vol. 1999, pp. 437-440, 1999.

[6] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, issue 2, pp.310-319, February 2010.

[7] J. Sundberg, *The Science of the Singing Voice*. DeKalb, IL: Northern Illinois University Press, 1987.

[8] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between singing and speaking voices," *Proc. Eurospeech*, vol. 2005, pp. 1141-1144, September 2005.

[9] M. Goto and T. Nishimura, "AIST Humming Database: Music database for singing research," *The Special Interest Group Notes of IPSJ (MUS)*, vol. 82, pp.7-12, 2005. (in Japanese)

[10] Y. Ohishi, M. Goto, K. Itou, and K. Takeda. "On the human capability and acoustic cues for discriminating the singing and the speaking voices," *Proc. Int. Conf. Music Percept. Cog.*, vol. 2006, pp. 1831-1837, August 2006.

[11] C. Wu and L. Gu. "Robust singing detection in speech/music discriminator design," *IEEE Int. Conf. Acoust. Speech Signal Process. 2001,* vol. 2, pp. 865-868, May 2001.

[12] B. Schuller, B. J. B. Schmitt, D. Arsic, S. Reiter, M. Lang, and G. Rigoll, "Feature selection and stacking for robust discrimination

of speech, monophonic singing, and polyphonic music," *IEEE Int. Conf. Multimedia Expo*, vol. 2005, pp.840-843, July 2005.

[13] T. Saitou, M. Goto, M. Unoki, and M. Akagi. "Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices," *IEEE Works. Appl. Signal Process. Audio Acoust.*, vol. 2007, pp. 215-218, October 2007.

[14] T. L. New, M. Dong, P. Chan, X. Wang, B. Ma, and H. Li, "Voice conversion: From spoken vowels to singing vowels," *IEEE Int. Conf. Multimedia Expo*, vol. 2010, pp.1421-1426, July 2010.

[15] S. Aso, T. Saitou, M. Goto, K. Itoyama, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno, "SpeakBySinging: Converting singing voices to speaking voices while retaining voice timbre," *Proc. Int. Conf. Digital Audio Effects (DAFx-10)*, vol. 2010, September 2010.

[16] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. "RWC music database: Popular, classical, and jazz music databases," *Proc. Int. Conf. Music Inform. Retrieval (ISMIR)*, vol. 2, pp. 287-288, October 2002.

[17] CMU Pronouncing Dictionary, www.speech.cs.cmu.edu/cgi-bin/cmudict.