Exploring Expressive Speech Space in an Audio-book

Lijuan Wang[2], Yong Zhao[1], Min Chu[1], Yining Chen[1], Frank Soong[1], Zhigang Cao[2]

[1]Microsoft Research Asia, Beijing, China
[2]Department of Electronic Engineering, Tsinghua University, Beijing, China
[1]{yzhao; minchu; ynchen; frankkps}@microsoft.com
[2]{wlj01}@mails.tsinghua.edu.cn, [2]{czg-dee}@tsinghua.edu.cn

Abstract

In this paper, an audio-book, in which a professional voice talent performs multiple characters, is exploited to investigate the expressiveness of speech. The expressive speech space of the sole speaker is explored by finding the distances between acoustic models of multiple characters and the perceived proximity between their speech utterances. Using the speech of ten characters as test data, the character confusion is evaluated in both acoustic space and perceptual space. We find that the average precision to differentiate one character from the others is 81.7% in the acoustic space and 72.6% in the perceptual space. It is interesting that the objective measure outperforms the subjective measure. Furthermore, the acoustic distance measured by normalized Kullback-Leibler divergence (NKLD) between two characters is highly correlated with the perceptual distance. The correlation coefficient is 0.814. Therefore, NKLD can measure the perceptual similarity between groups of utterances objectively.

1. Introduction

Synthesizing speech with rich expression is becoming an attractive topic recently in Text-to-Speech research. Contrasting to neutral speech spoken in a rather plain manner, expressive speech normally carries richer para-lingual and extra-lingual information that gives listeners cues on the emotional status, attitude, or intention of the speaker [1] [2]. In order to generate speech with a rich expression, we need to understand how people convey and perceive expressions in speech. In this paper, an audio-book is employed as a vehicle to study expressive speech. In the audio-book, a professional voice talent performs multiple characters in various expressions. The expressive space of the speaker is then explored by assessing the similarity of the acoustic models parametrized by Gaussian Mixture Models (GMMs) and the perceived space of the underlying characters.

When a person receives a phone call from a stranger, he normally portrays a virtual image of that stranger with his gender, age, emotion, mood, purpose, attitude or other characteristics, as illustrated in Fig. 1, according to the perceived speech. In such a process, at least three spaces can be used to characterize the underlying speech: the *expressive space*, which contains expressions planed consciously or unconsciously by the speaker; the *acoustic space*, which contains the physical signals generated by the speaker; and the *perceptual space,* which contains the expressions perceived by the receiver, as illustrated in Fig. 2.

The features used to describe the expressive space can be classified into two categories -- speaker or state related features. Speaker related features include gender, age, dialect, accent, tone-of-voice, and individual differences. State related features include emotion, speaking style, purpose, attitude, speaker-listener relationship, etc. The former is relatively steady within a speaker, while the later is controlled by the speaker. Comparing with neutral speech, they are the key that makes speech expressively. Each of these features can be one dimension in the *expressive space*.

Figure 1: Virtual images constructed by the receiver for a stranger during a phone call

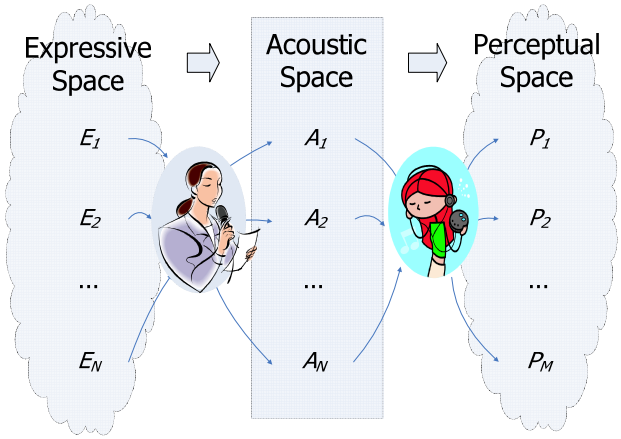

Fig 2: *Three spaces involved in expressive speech*

The features in the *perceptual space* could be the same as those used in the *expressive space*. Dissimilarly they are predicted from the acoustic signal of the perceived speech by the listener. The perceived expression is sometimes not the

same as the intended expression of the speaker because many factors, such as attention, emotion, past experience, expectation and environment of the receiver, will affect the listener's perception. In expressive speech synthesis, what we care most is how to adjust the acoustic features such that the speech perceived dose indeed convey the intended expressions. Therefore, the relationship between the acoustic space and the perceptual space is main focus of this study.

Collecting expressive speech data is crucial for studying expressions in speech. The data can be collected by the following ways [1]:

1. Hiring professional actors to speak with the specified expressions;
2. Inducing a subject to speak in a particular style or emotion by providing appropriate stimulus;
3. Selecting emotional speech segments from a large conversational speech database.

In the first two methods, the expressive speech collected can be overly acted or exaggerated. In the third method, the voice quality cannot be properly guaranteed and the expect emotion may not exist in the speech database.

This study uses a fiction audio-book narrated by a professional voice talent. Not only is the obtained speech of high-quality, but also the expressions in the full context of the audio-book sound more natural than just dubbing a single sentence with the assigned emotion. Containing multiple distinctive characters mimicked by one speaker is a unique feature of the audio-books in speech expressiveness. Since speech is the only means to present the whole story, the voice talent tries his best to perform different characters or the same character in different conditions by changing his sound. During the recording, he initially created a character list, then, matched each character to one of his neighbors or relatives. Therefore, he tried to mimic the voices of these acquaintances to distinguish different characters. The audio-book thus has a good coverage of the voice talent's expressive space.

As shown in Fig. 2, we assume that the voice talent tried to mimic N characters, represented as E_n (n=$1...N$). Then, A_n represents the acoustic sound he generated for the n-th character, and P_m ($m=1...M$) is the character identified by the listener in his/her perceptual space. M may not equal N because the acoustic signal may not provide enough discriminative information to the listeners. In this study, the discriminabilities among characters are examined in both the acoustic space and the perceptual space via character identification with acoustic models and a perceptual experiment, respectively. The relationship between the acoustic distances and perceptual distances among characters are investigated.

In the remaining paper, the data preparation is introduced in Section 2. The discriminabilities among characters in the acoustic space and the perceptual space are investigated in Section 3 and Section 4, respectively. The correlation between acoustic distances and perceptual distances is presented in Section 5. Section 6 draws the conclusions.

2. Data Preparation

The speech waveforms are first segmented into sentences aligned to the corresponding text script with the help of HTK toolkits [4]. Then, character identities are manually labeled for some sentences. There are dozens of characters appeared throughout this audio-book. We only select the top ten for studying, including the narrator. They are referred to as C01 ~ C10. 300 utterances are selected for each character. 250 of them are used as the training set and the other 50 as test set.

3. Character identification in acoustic space

Since all the characters are performed by the same speaker, it is interesting to first find out the discriminability among characters in the acoustic space. We employ state-of-the-art speaker identification technologies in the character identification task by assuming each character as an individual speaker. The performance of computer algorithms has been reported to be competitive with human listeners in speaker identificability (94.7%) [5]. However, in our case, discriminating different characters mimicked by the same speaker is much more challenging. We'd like to see whether computer still achieves competitive performance to human being.

3.1. Modeling the characters

In text-independent speaker recognition, a speaker is normally modeled with a Gaussian Mixture Models (GMM) [6]. A speaker independent GMM, or Universal Background Model (UBM), is first trained from a speech corpus of many speakers. Then, each target speaker's model is trained by adapting the UBM with utterances of the specific speaker in the Maximum A Priori (MAP) sense. In our study, UBM is trained from all utterances in the audio-book and each character model is adapted with 250 utterances of the character. During the adaptation process, only means are adjusted. MFCC's and fundamental frequencies and their delta coefficients are used as the acoustic features. Only voiced frames are used for training models because the vocal tract information of a specific speaker is embedded in voiced segments.

Once the character models are adapted, they are evaluated with a test set consisting of 50 utterances from each of the 10 characters. For each utterance in the test set, the acoustic likelihood measured against 10 character models and the UBM are calculated. The character whose model yields the highest likelihood is identified as the character. The precision of character identification is shown in Fig.3. The average precision is 81.7%, a performance much lower than that obtained in conventional speaker identification. From Fig. 3, it is seen that the identification rate for some characters, like C01, C06 and C07, are very high while for some others can be very low. In fact, some characters like C02 and C03 are highly confused with each other, an indication that the voice talent is not very successfully to distinguish the two characters well in his voice. It is desirable if we can find a more objective way to measure the similarity between different characters in the acoustic space.

3.2. Acoustic distance between characters

Since the voice characteristics of each character are captured by its character models, distance between two models should reflect the acoustic dissimilarity between the corresponding characters. Symmetric Kullback-Leibler divergence (KLD) between GMMs [7] is used in this study. Given a set of N character models, denoted as $\{\Lambda_n, 1 \le n \le N\}$, the symmetric KL divergence is defined as the sum of relative entropy

between model Λ_i and model Λ_j plus the relative entropy between model Λ_j and model Λ_i as shown in Eq. 1:

$$KLD_{(\Lambda_i,\Lambda_j)} = E_{\Lambda_i(X)}[\log\frac{\Lambda_i(X)}{\Lambda_j(X)}] + E_{\Lambda_j(X)}[\log\frac{\Lambda_j(X)}{\Lambda_i(X)}] \quad (1)$$

where $\Lambda_i(X)$ and $\Lambda_j(X)$ are the occurrences likelihoods of observation *X*, given Λ_i and Λ_j respectively. According to [7], normalized KLD (NKLD) in Eq. 2 fits human perception better. Therefore, the normalized KLD is used as the acoustic distance between two voice characters. By calculating the normalized KLD between each pair of characters, an *N*-by-*N* (N=10 in our case) symmetric acoustic distance matrix with zeros in the diagonal cells is obtained and shown in Table 1.

$$NKLD_{(\Lambda_i,\Lambda_j)} = \log(KLD_{(\Lambda_i,\Lambda_j)} + 1) \quad (2)$$

4. Character identification in perceptual space

Although we have measured the acoustic similarity between characters models by the normalized KLD, we are still not sure to what extent such a measure conforms to human perceptions. A subjective experiment was therefore carried out to measure the character distance in perceptual space.

4.1. The perceptual experiment

To make the task easy for subjects, utterances from the same character or different characters were paired and presented to the subjects. Subjects were asked to judge whether the two utterances are said by the same speaker or not (subjects didn't know that these utterances were in fact said by the same voice talent). 650 pairs of utterances were prepared for the experiment. Among them, 200 sentence pairs were inner-character comparison, 20 intra-character pairs for each character; 450 pairs are inter-character comparison, 10 pairs for between any two characters.

All utterance pairs were sorted randomly and separated into two sessions. Subjects were asked to finish the two parts with a not-less-than 30-minute break in between. The utterance pairs were played to the subjects with a scoring tool in a standard PC and subjects listened to them through headphones. The sequence of stimuli played to each subject was randomly generated. Subjects were allowed to listen to each pair as many times as they wanted before making the final decision of "same speaker" or "different speakers". After the choice is made, next utterance pair will be presented. On average, it took a subject 3 hours to finish the experiment. Before the formal testing, a short training session was carried out. 20 Chinese graduate students, fluent in English speaking and with normal hearing, participated in the experiment. No one had ever listened to this audio-book before. Also they don't know how many 'speakers' spoke in the experiment.

4.2. Perceptual distance between characters

To evaluate human identification accuracy in the perceptual space, we define a perceptual identification precision (PIP) for each character X in Eq. 3. $PIP_{(X)}$ is the number of correct decisions involving character X to the total number of pairs involving X. If a subject chose the "same" when paired utterances were from the same character or chose "different" when they were from different characters, the decision was regarded as a correct one. Otherwise, it was a wrong decision. The perceptual identification rate for the 10 characters is shown in Fig. 3. The average identification rate is 72.6%, which is much lower than that in acoustic space. Furthermore, the accuracy in perceptual space is more flat across characters than in acoustic space. This shows that the resolving power in the perceptual space is vaguer than in the acoustic space.

$$Perceptual\ \ Identification\ \ \Pr ecision_{(X)} = PIP_{(X)} \quad (3)$$

$$= \frac{number\ \ of\ \ correct\ \ decisions\ \ involving\ \ X}{Total\ \ number\ \ of\ \ pairs\ \ involving\ \ X}$$

The perceptual distance (PD) between two characters X and Y is defined by Eq. 4, i.e. the number of utterance pairs between X and Y were judged as "different" over the total number of pairs between X and Y. Small perceptual distance means character X and Y are perceptually similar and vice versa. The distances between each pair of characters are shown in Table 2.

$$Perceptual\ \ Dis\tan ce_{(X,Y)} = PD_{(X,Y)} \quad (4)$$

$$= \frac{number\ \ of\ \ pairs\ \ between\ \ X\ \ and\ \ Y\ \ judged\ \ as\ \ "different"}{number\ \ of\ \ pairs\ \ between\ \ X\ \ and\ \ Y}$$

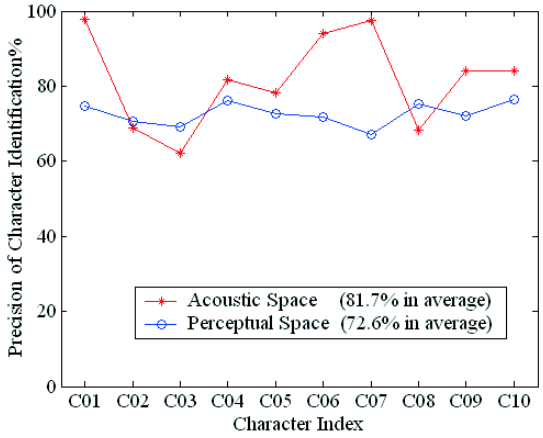

Fig 3: *The precision of character identification for the ten characters in the acoustic and perceptual spaces*

5. Relationship between acoustic distance and perceptual distance

To find out the relationship between the acoustic distance (Table.1) and the perceptual distance (Table.2), a scatter diagram for the two dimensions is plotted in Fig.4. The horizontal axis represents the acoustic distance between different characters and the vertical axis, the perceptual distance. The correlation coefficient between the two distances is 0.814, which indicates a high correlation between them. From this result, we can conclude that the acoustic distance measured by the normalized KLD between characters' GMMs correlates well with the subjective perceptual distance between these characters. Therefore, it can be used to objectively measure the perceptual similarity between groups of utterances.

Table 1: *Acoustic distance (NKLD) between different characters*

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10
C01	0.00	1.01	1.04	1.19	1.48	1.21	1.22	1.59	1.31	1.59
C02	-	0.00	0.66	0.85	1.17	1.10	0.97	1.35	1.24	1.44
C03	-	-	0.00	0.82	1.13	1.19	0.94	1.37	1.15	1.47
C04	-	-	-	0.00	1.20	1.37	1.07	1.29	1.42	1.61
C05	-	-	-	-	0.00	1.43	1.25	1.24	1.38	1.72
C06	-	-	-	-	-	0.00	1.09	1.25	1.26	1.32
C07	-	-	-	-	-	-	0.00	1.24	1.23	1.37
C08	-	-	-	-	-	-	-	0.00	1.53	1.60
C09	-	-	-	-	-	-	-	-	0.00	1.57
C10	-	-	-	-	-	-	-	-	-	0.00

Table 2: *Perceptual distance between different characters*

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10
C01	0.22	0.48	0.7	0.75	0.84	0.77	0.66	0.89	0.66	0.93
C02	-	0.44	0.61	0.71	0.81	0.74	0.7	0.91	0.68	0.87
C03	-	-	0.35	0.41	0.76	0.8	0.61	0.86	0.65	0.87
C04	-	-	-	0.32	0.85	0.9	0.7	0.88	0.86	0.88
C05	-	-	-	-	0.41	0.68	0.75	0.63	0.75	0.61
C06	-	-	-	-	-	0.48	0.68	0.71	0.75	0.63
C07	-	-	-	-	-	-	0.45	0.72	0.61	0.74
C08	-	-	-	-	-	-	-	0.51	0.83	0.61
C09	-	-	-	-	-	-	-	-	0.45	0.87
C10	-	-	-	-	-	-	-	-	-	0.35

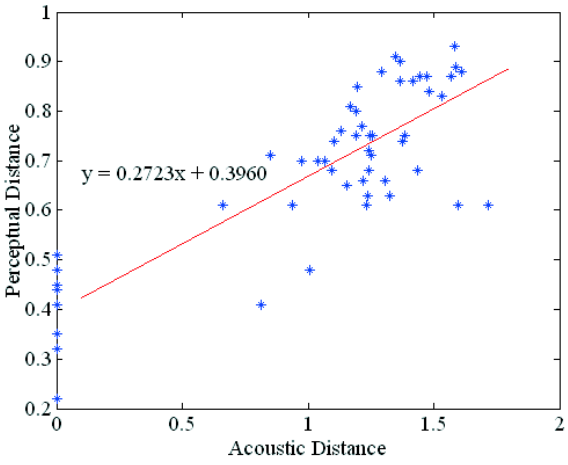

Fig 4: *Correlation and regression between acoustic and perceptual distance (The regression line is: y=0.2723x+0.3960.)*

Although the correlation between the two dimensions is rather high, there are inconsistent cases in the two tables. For example, C01 and C02, C03 and C04 are the closest pairs in the perceptual space; in the acoustic space, their distance is larger than many other pairs. The farthest character pair in the acoustic space is C05 and C10, yet their perceptual distance is not very large. While listening to these characters, we have found that the voice talent uses many ways to distinguish them. He increases or decreases pitch register for some characters and changes speech rate for others. He also adds special accent to some characters. Some of these changes, such as the change in pitch can be captured by the current character acoustic model but others like speech rate or accent cannot. As a result, some characters distant in the perceptual space are close in the acoustic space. On the other hand, since all characters are performed by the same speaker, the listener sometimes is not sensitive to the subtle timbre variation. As a result, some characters distant in the acoustic space are closer in the perceptual space.

6. Conclusions

In this work, an audio-book is employed as an expressive speech database to investigate the expressiveness of speech, which contains multiple characters performed by a professional voice talent. The expressive speech space of the voice talent is explored by assessing the distances between acoustic models of multiple characters and the perceived proximity between the corresponding speech utterances. Using the speech of ten characters as test data, the character confusion is evaluated in both acoustic space and perceptual space. The acoustic distance and perceptual distance between any two given characters are calculated and their correlation is computed. We find that the average precision to differentiate one character from the others is 81.7% in the acoustic space and 72.6% in the perceptual space. It is interesting that the objective measure outperforms the subjective measure. Furthermore, the acoustic distance measured by normalized Kullback-Leibler divergence (NKLD) between two characters is highly correlated with the corresponding perceptual distance measured subjectively by human listeners. The correlation coefficient is 0.814. Therefore, the acoustic distance can be used as an objective measure for the perceptual similarity in term of expressiveness between groups of utterances. In future study, we will work on utterance clustering from the speech expressiveness perspective.

7. References

[1] M. Tatham and K. Morton, “Expressive in Speech: Analysis and Synthesis,” Oxford university press, 2004.

[2] N. Campbell, “Labeling natural conversational speech data,” 1-10-22, 273-4 in Proceeding of ASJ Fall meeting, 2002.

[3] H. Traunmuller, “Perception of speaker sex, age, and vocal effort,” http://www.ling.su.se/staff/hartmut/F97.pdf, 1997.

[4] J. Odell, D. Ollason, P. Woodland, S. Young and J. Jansen, The HTK Book for HTK V3.0, Cambridge University Press, Cambridge, UK, 2001.

[5] A. Schmidt-Nielsen and T. H. Crystal, “Human vs. Machine Speaker Identification with Telephone Speech,” in Proceeding of ICSLP, 1998.

[6] X. D. Huang, A. Acero, H. Hon, Spoken Language Processing, Prentice Hall, New Jersey, 2001.

[7] S. Kullback and R. Leibler, “On information and sufficiency,”Annals of Mathematical Statistics, vol. 22, pp. 79–86, 1951.

[8] M. Sakamoto and T. Saito, “Speaker Recognizability Evaluation of a Voicefont-based Text-to-speech system,” in Proceeding of ICSLP, 2002.

[9] P. E. Papamichalis and G. R. Doddington, “A Speaker Recognizability Test,” in Proceeding of ICASSP, 1984.