# VOICE QUALITY: NOT ONLY ABOUT "YOU" BUT ALSO ABOUT "YOUR INTERLOCUTOR"

*Ya Li[1,2], Nick Campbell[3], Jianhua Tao[1]*

[1] National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] CNGL, Trinity College Dublin, the University of Dublin, Ireland
[3]Speech Communication Lab, Centre for Language and Communication Studies, Trinity College Dublin, the University of Dublin, Ireland
yli@nlpr.ia.ac.cn, nick@tcd.ie, jhtao@nlpr.ia.ac.cn

## ABSTRACT

This paper investigates the effect of voice quality in commutative speech. Voice quality is often considered as the characteristic auditory colouring of an individual speaker's voice, but in our study, we find that voice quality can also reveal information about the interlocutor in everyday social interactions. In the correlation analysis between acoustic measures and interlocutors, the effect caused by the linguistic content was reduced by focusing on the corpus of the commonly-used Japanese word "yes". The distributions of voice quality features, e.g., Normalized Amplitude Quotient (NAQ), jitter and shimmer, showed a clear difference among different interlocutors, e.g., friend or business partner. Automatic classification of interlocutors was conducted using random forest method with voice quality, prosodic and spectral features. The best classification accuracy was 76.9% over four interlocutors' data.

***Index Terms***— Voice quality, prosody, decision trees, commutative speech, social signal

## 1. INTRODUCTION

Speech carries more than text content. When we speak, our mood, attitude and affect are also encoded in the speech signal mainly by means of prosody [1]. Prosody includes the rhythm, stress, and intonation of speech, which is used to convey social and paralinguistic information, and usually signaled by changes in the pitch, duration and energy. But, in a recent study, voice quality is considered as the 4th dimension of prosody [2] in much the same way as, for example, pitch. A similar statement is that voice quality and prosodic features are all considered as part of the vocal social signals which affect the perception of a message [3]. Voice quality, in a broad sense, is described as [4] *"the characteristic auditory colouring of an individual speaker's voice, and not in the more narrow sense of the quality deriving solely from laryngeal activity. Both laryngeal and supralaryngeal features will be seen as contributing to voice quality."* The variation of voice quality could reveal the speaker's emotional state [5], and attitude [5] towards the listener or the content or both, and also the speaker's social status [6], cultural background [7] and personal characteristics [8]. For instance, Gobl *et al.* [5] explored the role of voice quality in communicating emotion, mood and attitude based on synthetic stimuli, and they found that voice quality has more impact on milder emotions than that on strong emotions. More recently, Charfuelan *et al.* [6] used scenario meetings corpus to detect the social status of the attendees, and they found the most dominant person tends to speak with a louder-than-average voice and the least dominant person with a softer voice.

Different from the previous research, the intuition of this study is that from people's voice we are not only able to tell their affect, mood and attitude, but we can also make a fair guess of who they are interacting with, even when the interlocutor is invisible in a telephone conversation. For example, when people speak to their family, they tend to use a soft voice; on the contrary, a formal and gentle voice is the common voice when people speak to their boss. These observations motivate us to ask whether the interlocutor's information is encoded in communicative speech or not. If yes, how is it encoded? This paper attempts to answer these questions by analyzing a conversational speech corpus of everyday social interactions. Here the difference in the linguistic content of the speech was reduced by focusing on the analysis of the commonly-used Japanese word "yes". The correlation analysis between several voice quality measures and prosodic features and different interlocutors were carried out. Finally, we also used these features to classify interlocutors automatically by random forest method to verify this assumption. These experiments showed that the information about speaker's interlocutor is encoded in their speech.

The paper is organized as follows. Section 2 introduces some measures of voice quality and prosody features which

were used in this work. In Section 3, the corpus and data analysis are presented. Section 4 will give the details of automatic interlocutor classification with the voice quality, prosodic and spectral features. Conclusions and main findings are summarized in Section 5.

## 2. VOICE QUALITY

There are several acoustic measures of Voice Quality (VQ). The VQ measures used in this work are introduced below.

Jitter and shimmer can be defined as the variations of fundamental frequency and amplitude between consecutive periods, respectively [9]. High jitter levels often indicate that normal vocal fold vibration is interfered, and the speech signal is usually classified as roughness and/or breathiness.

The difference between the first and second harmonic amplitudes (H1-H2) is a spectral measure and used to describe the harmonic structure. H1-H2 is an effective cue for detecting creaky voice. Another spectral feature used in this work is the Harmonic to Noise Ratio (HNR) [10], which represents the degree of periodicity. This parameter is defined as the ratio of power between the periodic and aperiodic components of the speech.

There are other VQ measures obtained from the glottal flow estimated from speech signal [11]. Amplitude Quotient (AQ) is the ratio between glottal signal amplitude and the minimum value of the glottal signal derivative, while Normalized Amplitude Quotient (NAQ) is the normalized form of AQ based on fundamental frequency, which is more robust than AQ [12]. It is also suggested that NAQ is an effective parameter for separating breathy to tense voice [12].

## 3. VOICE QUALITY ANALYSIS ON CONVERSATIONAL SPEECH

### 3.1. Corpus

A large amount (1500 hours) of fluent conversational speech with English, Chinese, and Japanese interlocutors of everyday social interactions has been collected as part of the JST/CREST Expressive Speech Processing (ESP) project [13]. The speech conversations of speaker FAN's data was chosen to verify the assumption. FAN is a female Japanese speaker, and she wore a small head-mounted, studio-quality microphone everyday throughout the data collection to record the real spoken interactions. The interlocutors ID of her conversation were manually labeled.

In order to reduce the effect of linguistic content on correlation analysis between speech signal and interlocutors, we focused on one commonly-used example; the word 'hai' (a Japanese expression which means 'yes' in English). This word is widely used in our daily life to express different meanings, affects and attitude. The utterances of the single "hai" were firstly selected from the 600-hr FAN subset of the same ESP corpus according to their time stamp labeling automatically. In order to reduce the effect of the linguistic content and also simplify this work, complex utterances, such as "hai hai" "hai hai hai" were not considered. The average duration of the selected segments is very short (365.5 ms). After data selection, there are a total of eight interlocutors in the corpus, which is divided into main part and secondary part according to their number of instances. The distribution of utterances among the different interlocutors is indicated in Table 1.

Table 1. Interlocutors distribution in the subset of the extracted utterances

| | Label in the corpus | Interlocutor | # of instances |
|---|---|---|---|
| Main part | ane | older sister | 52 |
| | shiyakusho | city hall | 66 |
| | tomodachi | friends | 300 |
| | yubinkyoku | post office | 22 |
| Secondary part | biyouin | beauty shop | 7 |
| | haken | employer | 6 |
| | ntt | telephone company | 2 |
| | oi | cousin | 6 |

### 3.2. Feature extraction

Besides the VQ measures, the most common used features in related research, such as: fundamental frequency (F0), intensity, duration and voicing probability were also utilized. Jitter, shimmer, HNR, pitch, duration, energy and spectral parameters were extracted using OpenSMILE [14] with the *avec2011.conf* config file. Apart from these low-level descriptors, OpenSMILE also uses statistical functions to generate high-dimensional feature vectors from these low-level features, such as, maximum, minimum, mean, standard variation and regression coefficients. The NAQ [12] and H1-H2 were extracted by the Voice analysis toolkit provided by [15]. It is reported that the algorithms proposed in [15] are more robust and accurate at estimating glottal source signals. The maximum, minimum, mean, and standard variation of NAQ were also calculated.

### 3.3. Correlation analysis

Figure 1 shows H1-H2 plotted against NAQ, for the different interlocutors. We notice that NAQ correlates well with 'the relationship between FAN and the interlocutor' or 'formality of the conversation'. This hypothesis is supported by observing that a clear groupings of familiar interlocutors, e.g., *friends*, *cousin* and *older sister*, which are clustered together in the lower left corner. In contrast, the three business partners: *beauty-shop*, *employer* and *telephone company* are grouped in the higher right corner. In this figure, *post office* is closer to familiar interlocutors, which imply that she treat the staff who work in post office more like her friend. The *city hall* falls into intermediate position,

relatively closer to the group of business partners, which perhaps due to a non-commercial relationship between FAN and this interlocutor. Fig. 4 indicates that higher NAQ is used when addressing other people politely [2].

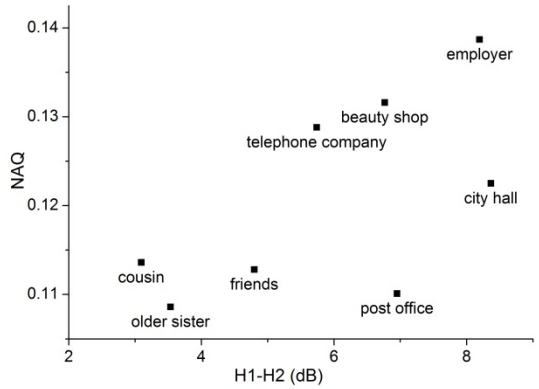


Fig. 1 H1-H2 plotted against Normalized Amplitude Quotient (NAQ) for the different interlocutors.

Figure 2 shows F0 plotted against NAQ for speaker FAN addressing different interlocutors. It can be seen that FAN generally speaks to *telephone company* with very high F0 value. High pitch is preferred for Japanese woman to express their politeness and femininity [16]. High NAQ (breathiness) and high F0 indicate how "careful" does the speaker talk to different interlocutors [2], which could be seen from the trend along the diagonal line of Fig. 2. There is no clear grouping associated with F0. This figure also suggests that NAQ provides more information about the interlocutors than F0.

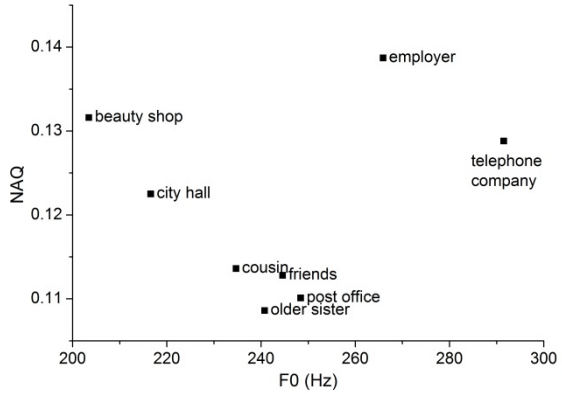


Fig. 2 Fundamental Frequency (F0) plotted against NAQ for the different interlocutors.

Figure 3 shows the distribution of jitter and shimmer for different interlocutors. Clear groupings can be seen: *friends* and *older sister* are associated with high jitter and shimmer, while *post office*, *employer* and *city hall* are associated with low jitter and shimmer. Jitter and shimmer are related to the irregularity and perturbation of the pitch and amplitude respectively [17]. We can expect that people tend to show their affect status to their friends and family, and they also tend to be involved more with their friends and family, and more relax, hence less formal. In such conversation context, more variation may occur in their neuromuscular control, and then the vocal folds, and finally lead to higher jitter and shimmer in their speech. On the contrary, when people speak to their business partners, they tend to modulate their speech, which is reflected on lower jitter and shimmer.

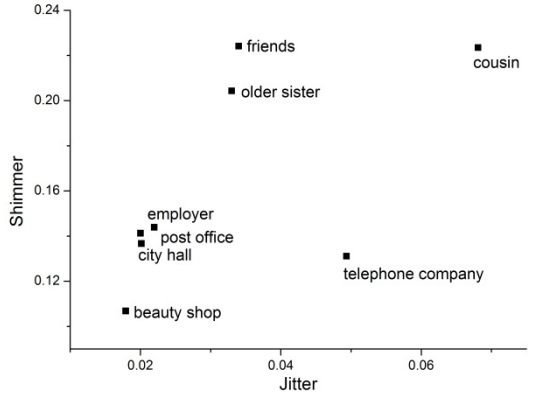


Fig. 3 Jitter and shimmer distribution for the different interlocutors.

HNR is to measure the additive noise in speech, and often used to measure the degree of breathy and hoarse voice [10]. Fig. 4 shows the distribution of HNR against F0 for the different interlocutors. Fig. 4 indicates that HNR is higher for interlocutors (except *telephone company*) who have a closer relationship with the speaker and it also shows that F0 is a better cue to distinguish *post office*, *city hall* and *employer*, which could not in Fig. 3. This is explained by the fact that people instinctively lower their pitch and energy when they speak to people of a higher social status, e.g., city hall officer, to make their voice less aggressive [18].

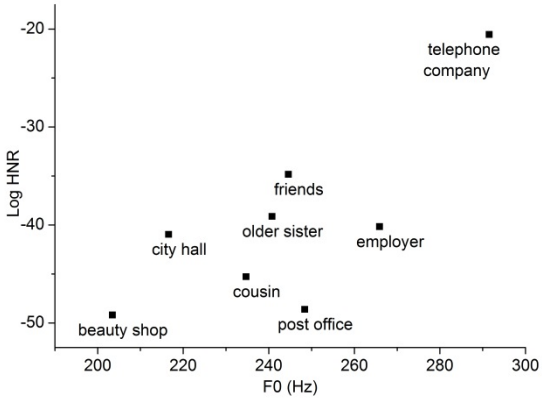


Fig. 4 Fundamental Frequency (F0) plotted against Harmonic to Noise Ratio (HNR) for the different interlocutors.

## 4. INTERLOCUTOR CLASSIFITION

In the previous section, results of the acoustic measurement analysis indicated that voice quality provides information about the interlocutor, e.g., their relationship and their social status with the speaker. In order to test if it is possible to automatically predict the speaker's interlocutor from the speaker's recordings, we developed an automatic interlocutor classifier using machine learning approach.

Only the main part of the corpus, which is listed in Table 1, is considered for classification. The secondary part of the corpus was removed because their instances are few to train a classifier, and some instances of the main part were also removed due to the difficulty in extracting part of the features.

Two feature sets were designed for comparison. The first one used all the features obtained using OpenSMILE and Voice analysis toolkit provided by [15]. In total, it consists of 1947 features which can be roughly divided into voice quality, prosodic and spectral related features. These spectral features mainly include features based on LSP and MFCC. On the other hand, the second feature set includes the voice quality and prosodic features.

The machine learning method used in this study was Random forest, and we used the *weka* implementation of this method [19]. Random forest is a combination of multitude decision trees, in which the final prediction is made by aggregation of the majority vote or averaging the each separate decision tree. The number of trees used in this work was ten. In the preliminary experiments, other classification methods (SVM, decision tree, RBFnetwork, and Naive Bayes) have been tested too, and random forest obtained the highest accuracy, which was approximately 2-10% higher than the others. Ten-fold cross-validation was conducted to assess the classifier. Table 2 shows the overall classification results of the Random Forest method for the two feature sets. Since the distribution of the classes is not balanced, unweighted accuracy (an average of the per-class accuracies) is a more reasonable measure to assess the results. Results also show that it is possible to obtain fair classification accuracies using voice quality and prosodic features alone. These results are promising given that only acoustic measures were used in this work. The results could be further improved by considering linguistic content of the speech as well. For example, some certain words, *darling* and *professor* could strongly indicate people's interlocutor.

Table 2. Interlocutor classification results using random forest (WA= weighted accuracy, UA= unweighted accuracy)

| Feature set | WA (%) | UA (%) |
|---|---|---|
| 1 (Prosody+VQ+Spectra) | 76.9 | 67.1 |
| 2 (Prosody+VQ) | 65.4 | 48.8 |

Table 3 shows the confusion matrix of the classification results. It shows that *friends* and *older sister* are easy confused with each other by the classifier. 50.0% of the *older sister* is classified as *friends,* and 11.6% of *friends* are classified as *older sister*. The confusion between *city hall* and *post office* are 11.3% and 6.1% for misclassifying one as the other, respectively. This agrees with the voice quality analysis in the previous section, because *friends* and *older sister* have similar values of voice quality measures, and they seemed to be grouped together, while *city hall* and *post office* usually grouped together. In other words, the speaker tends to have same similar spoken interaction with people who belongs to the same social group. This effect has already been observed for social signals in general: verbal and non-verbal [3].

Table 3. Confusion matrix of the classification results

| Annotated / Classified | older sister | city hall | friends | post office |
|---|---|---|---|---|
| older sister | 8 | 2 | 42 | 0 |
| city hall | 0 | 45 | 22 | 1 |
| friends | 8 | 7 | 285 | 8 |
| post office | 0 | 8 | 12 | 2 |

## 5. CONCLUSIONS

This paper presents our contribution to investigate the social information encoded in the communicative speech other than linguistic content. Several voice quality measures and prosodic features were extracted and analyzed from a subset of a large Japanese daily life recording. We found that both voice quality and prosodic features shown to be correlated with the information about the speaker's interlocutor, e.g., relationship and social status difference between them. This work extends previous research which mainly focused on the correlation between voice quality and speaker's characteristics. Automatic interlocutor classification was also carried out to verify our assumption of social correlates. Experiments showed that promising interlocutor classification accuracy can be obtained from a set of prosodic and voice quality features alone. The validation on a larger scale corpus will be carried out later. Future work includes considering both of the speaker's characteristics and the interlocutor information together into analysis of the para-linguistic content of social prosody in communicative speech.

## 6. ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC) (No.61203258, No.61305003, No.61375027), the Major Program for the National Social Science Fund of China (13&ZD189) and CNGL (12/CE/I2267) project, and partly supported by the Open Projects Program of National Laboratory of Pattern Recognition (201407353).

## 7. REFERENCES

[1] H. Fujisaki, "Information, Prosody, and Modeling," *Proceedings of Speech Prosody, Nara, Japonia,* 2004.

[2] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *15th ICPhS*, pp. 2417-2420, 2003,.

[3] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico*, et al.*, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *IEEE Transactions on Affective Computing,* vol. 3, pp. 69-87, 2012.

[4] J. Laver, "The phonetic description of voice quality," *Cambridge Studies in Linguistics London,* vol. 31, pp. 1-186, 1980.

[5] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication,* vol. 40, pp. 189-212, 2003.
[6] M. Charfuelan, M. Schröder, and I. Steiner, "Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings," in *INTERSPEECH*, pp. 2558-2561, 2010.
[7] M. O. Sven Grawunder, Cordula Schwarze, "Politeness, culture, and speaking task – paralinguistic prosodic behavior of speakers from Austria and Germany," in *Speech Prosody*, pp. 159-163, 2014.
[8] N. Campbell, "Listening between the lines: a study of paralinguistic information carried by tone-of-voice," in *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*, 2004.
[9] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *INTERSPEECH*, pp. 778-781, 2007.
[10] E. Yumoto, W. J. Gould, and T. Baer, "Harmonics-to-noise ratio as an index of the degree of hoarseness," *Journal of The Acoustical Society of America,* vol. 71, pp. 1544-1550, 1982.
[11] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis", *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, pp. 195-208, 2008.
[12] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parameterization of the glottal flow," *the Journal of the Acoustical Society of America,* vol. 112, pp. 701-710, 2002.
[13] N. Campbell, "Building a Corpus of Natural Speech–and Tools for the Processing of Expressive Speech–the JST CREST ESP Project," in *Proc. 7th European Conference on Speech Communication and Technology. Center for Personkommunikation (CPK), Aalborg*, pp. 1525-1528, 2001.
[14] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in openSMILE, the munich open-source multimedia feature extractor," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 835-838, 2013.
[15] J. Kane, "Tools for analysing the voice Developments in glottal source and voice quality," Doctor of Philosophy, Trinity College Dublin, 2012.
[16] L. Loveday, "Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae," *Language and Speech,* vol. 24, pp. 71-89, 1981.
[17] I. R. Titze, "Toward standards in acoustic analysis of voice," *Journal of Voice,* vol. 8, pp. 1-7, 1994.
[18] M. Ito, "Politeness and voice quality-the alternative method to measure aspiration noise," in *Speech Prosody,* 2004.
[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations,* vol. 11, 2009.