行政院國家科學委員會專題研究計畫 成果報告

以時、頻域多重解析聽覺模型為基礎之客觀語音品質估測(II)
研究成果報告(精簡版)

計畫類別：個別型
計畫編號：NSC 97-2221-E-009-114-
執行期間：97年08月01日至98年07月31日
執行單位：國立交通大學電信工程學系（所）

計畫主持人：冀泰石

計畫參與人員：碩士班研究生-兼任助理人員：葉藍霙
碩士班研究生-兼任助理人員：林廷翰
碩士班研究生-兼任助理人員：楊禮瑋
碩士班研究生-兼任助理人員：周文勝
博士班研究生-兼任助理人員：林澤恩

報告附件：出席國際會議研究心得報告及發表論文

處理方式：本計畫可公開查詢

中 華 民 國 98年10月21日

PERCEPTION-BASED OBJECTIVE SPEECH QUALITY ASSESSMENT

Ting-Yu Yen, Jian-Hueng Chen† and Tai-Shih Chi**

*Department of Communication Engineering
National Chiao Tung University, Taiwan

†Multimedia Applications Laboratory
Chunghwa Telecom Co., Ltd., Taiwan

ABSTRACT

A joint spectro-temporal auditory model is utilized to assess speech quality objectively. The model mimics early and central auditory functions and serves as a spectro-temporal modulation filterbank. Three perceptual relevant parameters, intelligibility, clarity and naturalness, are addressed by the model and are combined to estimate the subjective mean opinion score (MOS) for speech quality measure. Through a simple multiple linear regression analysis, we demonstrate the performance of our proposed perception-based objective speech quality measure is better than that of the state-of-the-art P.563 standard in estimating MOS of the codec-distorted speech in ITU-T Supp. 23 database.

Index Terms**—** objective speech quality, intelligibility, clarity, naturalness, auditory representation

1. INTRODUCTION

Speech quality often represents an end user's direct perception on the quality of a voice communication network. Over the past decades, more and more complex telecommunications networks were built to be involved in people's modern daily lives significantly. Nowadays, customers demand the quality of service from voice communication networks as distinct from only the connectivity at early days. Therefore, assessing speech quality effectively becomes an important task for voice service providers. In addition, a speech quality meter can also serve as a diagnostic tool to isolate damaged components in a network for the purposes of online quality monitoring or offline system maintenance.

There are two classes of speech quality measures: subjective and objective. Subjective methods ask a panel of listeners to rate the quality of tested speech signals. In general, listeners are asked to rate speech quality in a five-point scale (1: bad; 2: poor; 3: fair; 4: good; 5: excellent) and the average of all listeners' scores is called Mean Opinion Score (MOS). Although conducting subjective listening tests is the most reliable and natural way to assess speech quality, it is time and money consuming with a score which is almost impossible to reproduce. Therefore, it is crucial to develop computational objective methods to estimate subjective MOS with high fidelity, hence, to replace subjective listening tests.

Objective methods are further categorized into intrusive (double-ended) or non-intrusive (single-ended), based on the availability of the reference (original clean) speech. Intrusive methods measure the difference between the degraded and the original clean speech. PESQ, standardized in 2001, is the current state-of-the-art intrusive speech quality measure [1]. On the other hand, non-intrusive algorithms evaluate speech quality without the knowledge of the reference speech. The most common element in non-intrusive algorithms is an artificial reference model of the clean speech, either in VQ (Vector Quantization), HMM (Hidden Markov Model) or GMM (Gaussian Mixture Model) parameter spaces (see [2] for more detailed introduction). Then, the degraded signal is compared to the artificial clean template to estimate the MOS. P.563, standardized in 2004, is the state-of-the-art non-intrusive algorithm at present [3]. However, these algorithms are all built from the viewpoint of signals, but not from the viewpoint of perception.

In our opinion, speech quality is a multi-dimensional percept, which consists of at least three abstract percepts: intelligibility, clarity and naturalness. In this study, we address these three percepts separately. First, we utilize an auditory model to extract joint spectro-temporal modulations of speech signals [4]. In our model, these joint modulations are perceived by human brains and certain ranges of modulations are presumed related to these three percepts. A brief description of the model is given is section 2. In section 3, we extend the measurements proposed in [5][6], which have been shown to successfully estimate speech intelligibility under various noise conditions, to assess clarity and naturalness. Finally, a linear regression analysis is used to combine measures from these three percepts to match the subjective MOS. Our approach yields a better estimation of MOS for speech samples in ITU-T Supp. 23 database [7] than P. 563, as demonstrated in section 4.

2. AUDITORY MODEL AND REPRESENTATIONS

The auditory model used in this study is based on biophysical, psycho-acoustical and neuro-physiological evidences discovered in the stages of early cochlea and central auditory cortex. It consists of two computational modules which transform the time waveform into an *auditory spectrogram* (a time-logarithmic frequency pattern of the auditory nerve activity), then into a *scale-rate representation* (a pattern of the cortical neural activity).

The first module mimics observed functions of the peripheral auditory processing. As shown in the Figure 2 of [4], this module is implemented by a bank of 128 overlapping asymmetric constant-Q bandpass filters equally spaced on the log-frequency axis covering a total of 5.3 octaves (which simulates the frequency selectivity of the cochlea), followed by a non-linear compression and a lateral inhibitory network (which are pertaining to well-known psycho-acoustical properties of amplitude compression and frequency masking), and ended with an envelope extractor (which simulates the further leakage of current and the reduced temporal sensitivity of the auditory nerve). In this study, we use a simplified linear version of this module. The output of this first module represents a time-frequency energy distribution along a log-frequency axis, in a very rough sense similar to the traditional spectrogram plotted along the bark or mel scale.

The second module mimics the spectro-temporal selectivity of neurons in the primary auditory cortex (A1). Briefly, the output of the previous module, the auditory spectrogram, is then analyzed by a family of neurons which are modeled by two-dimensional filters tuned to different spectro-temporal modulation parameters. One of the parameter is referred to as **scale** or density (in cycle/octave), which defines how broad the signal energy distributed along the log. frequency axis. Another parameters is called **rate** or velocity (in hertz), which defines how fast the signal energy varies along the temporal axis. In addition to the scale and rate, directional selectivity for FM sweep of cortical neurons is also encoded in this module and is represented by the sign of the rate (negative for upward sweeping; positive for downward sweeping). In other words, this module performs a joint spectro-temporal multi-resolution analysis on the input auditory spectrogram. Fig.1. demonstrates the multi-resolution analysis by eight neurons which tune to different combinations of scale and rate. The top panel shows the input auditory spectrogram from the first module. The eight larger panels at the bottom are the outputs of the cortical neurons, whose impulse responses are depicted in the eight smaller panels respectively. We then compute the average excited power of each neuron by collapsing its two-dimensional output into a single value. Such diagram of average excited power, labeled by each neuron's {scale, rate} tuning characteristic, is referred to as the scale-rate representation. Much more extensive details of the description, mathematic formulation and output examples of the model can be found in [4].

3. PERCEPTUAL PARAMETERS AND SPEECH QUALITY ASSESSMENT

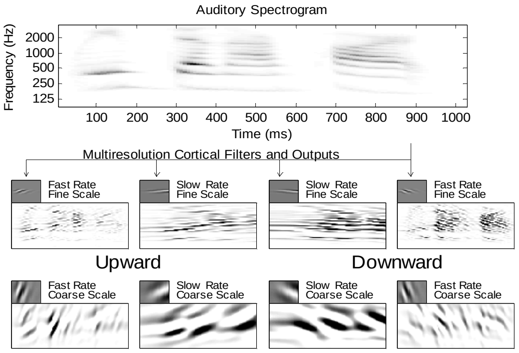

Fig. 1. Multi-resolution analysis by the auditory cortex.

The scale-rate representation mentioned in section 2 has been used to quantify the speech intelligibility with great success [5][6]. In this paper, we extend the measurement to assess another two important percepts: clarity and naturalness, and combine three percepts to estimate the speech quality.

3.1. Intelligibility

Intelligibility is defined by the ISO 9921 standard as a measure of effectiveness of understanding speech. Generally speaking, intelligibility refers to "what" a speaker has said, while speech quality refers to "how" an utterance was spoke. Usually, unintelligible speech is judged to be low quality, however, the converse needs not to be true.

We have shown that the spectro-temporal modulations between 4~8 Hz in rate and < 4 cycle/octave in scale are dominated in speech signals [6]. These very slow temporal modulations reflect the speed of the articulatory motions, and hence the phonetic and syllabic rates of speech [4]. Therefore, these slow temporal modulations are critical for speech intelligibility (or speech recognition) as suggested in [8]. We have further shown that speech reconstructed from scale-rate modulations up to 32 Hz and 4 cycle/octave does not suffer from loss of intelligibility [4]. To sum up, the modulations below 32 Hz in rate and 4cycle/octave in scale are measured to quantify the speech intelligibility in this paper.

3.2. Clarity

Although reconstructed speech from only low rate and low scale modulations well preserves speech intelligibility [4], listeners often report a "dull" sensation. On the other hand, original speech with full ranges of modulations gives a relatively "bright" sensation. Therefore, we assume that modulations of high rates and high scales, which provide more fine structures of the speech [4], would enhance speech clarity. Here, we use joint spectro-temporal modulations (32~128 Hz; 2~8 cycle/octave) to measure the

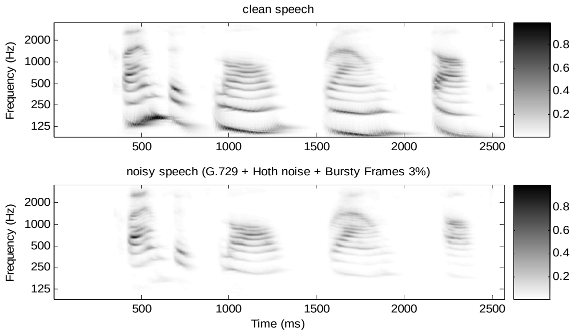

Fig. 2. Auditory spectrograms of samples from ITU-T Supp. 23 database (Top: clean speech; bottom: degraded speech after codec and noise deterioration).

speech clarity. Our assigned ranges cover the pure temporal modulation (30~50 Hz) used in [9] for speech quality measure.

From the viewpoint of speech processing, temporal filters with higher rate than 32 Hz do pick up important features related to the clarity of speech, such as onset/offset. The usual 16 ms frame duration used in a traditional frame-by-frame speech processing scheme is right within the temporal region we chose for the clarity percept (32~128 Hz). Meanwhile, the 2~8 cycle/octave provide more detailed spectral shapes (in a rough sense similar to the effect of using more cepstral coefficients in traditional analysis), which convey characteristics of consonants. It is believed that consonants/noise give the "bright" sensation and account for the clarity of speech.

3.3. Naturalness

Naturalness was defined by Parrish back in 1951 as speech that sounds natural or normal to the listener. In Parrish's concept of naturalness, more natural speech helps the listener to pay more attention to the meaning of words rather than to the speech pattern used to convey the meaning [10]. In other words, unconventional speech styles deteriorate the naturalness of speech. In 1978, Sanders et al. investigated possible factors of the naturalness for synthesized speech, such as pitch, duration, loudness and spectral contour [11]. Furthermore, studies also conclude judging the naturalness of speech does not necessarily require the same resources as judging intelligibility [12], which validate our approach of assessing intelligibility and naturalness separately.

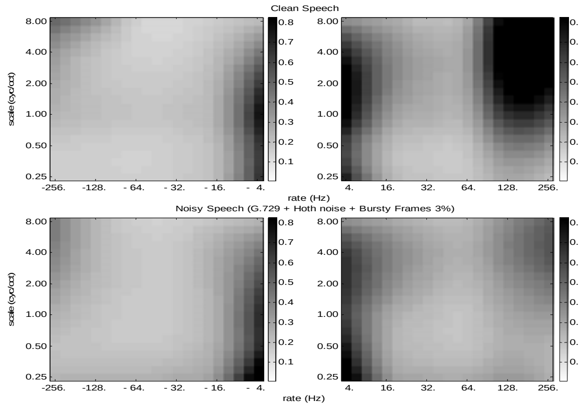

Fig. 3. Scale-rate representations speech samples in Fig. 2.

Fig. 2 and Fig.3 demonstrate auditory representations of speech samples from ITU-T Supp. 23 database (oe3m4519). The noisy speech shown at the bottom is degraded through a G.729 speech codec, hoth noise and 3% frame drop. Obviously shown in Fig. 2, the speech rate does not change noticeably even with 3% frame drop, while the fundamental frequency (F0) contour is missing after the degradation. This fact is even more highlighted in Fig. 3, where the strong energy distribution between 128~256 Hz (due to the F0) in the top panels almost disappears in the bottom panels. All degraded speech samples through a codec or cascades of codecs in the database have similar outcomes. Therefore, we conclude that the pitch dispersion is the most prominent degradation through speech codecs and is crucial to the loss of naturalness. In this paper, we use the joint spectro-temporal modulations (128~512 Hz; 0.25~8 cyc/oct) to quantify the degradation of naturalness.

3.4. Measurement of degradations

We have developed Spectro-Temporal Modulation Indexes (STMI) in [5][6] to measure speech intelligibility under a wide range of noisy environments. Here, we extend the same measurements (STMI^T in [5] and ρ in [6]) to assess clarity and naturalness percepts. Both measurements generate very similar results. All results shown in next section are calculated from ρ, which is defined as follows:

$$\rho = \overline{\left[1 + \left(\frac{T - N}{N_{\mathrm{std}}}\right)^2\right]^{-1/2}}$$

where T and N are the scale-rate representations of a clean template and the degraded testing speech, respectively; N_{std} is the standard deviation of the testing speech' scale-rate representation.

4. RESULTS

The long-term averaged scale-rate template of clean speech T is calculated from all 4620 sentences in the training portion of TIMIT corpus. Obviously, this template does not contain useful information within the naturalness range since the pitch values are averaged over 462 speakers. We add white noises with various SNRs (-15~45 dB) into sentences from the testing portion of TIMIT corpus and compute their ρ of intelligibility and clarity, respectively. Fig. 4 shows results plotted as functions of the estimated

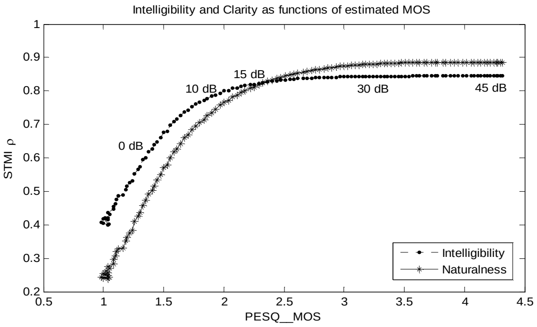

Fig. 4. Intelligibility and clarity as functions of estimated MOS by PESQ under AWGN conditions.

MOS by PESQ. It shows the intelligibility and clarity saturate around MOS:2.5 and MOS:3.5, respectively. These results clearly show the intelligibility and clarity are only partial aspect of speech quality. Moreover, these results are consistent with our informal listening tests that sentences with PESQ_MOS: 2.5 are all intelligible (recognizable), and are with higher and higher clarity when PESQ_MOS increasing to 4.5. The ρ does not approach to the unity in Fig. 4 due to the fact that we use a generic template (i.e., template and testing speech are not from the same speaker). See [5] for more discussions on using a generic template.

The long-term template of naturalness (128~512 Hz; 0.25~8 cyc/oct) of a speaker is generated by averaging his speech samples. From our observations, all long-term templates of naturalness are alike among different speakers except the shift on the rate axis due to different pitches. The ρ of three percepts: intelligibility, clarity and naturalness are each mapped through a sigmoid function and then are combined by using a multiple linear regression to estimate the subjective MOS. Our speech quality estimate is referred to as PB_MOS (Perception-Based estimate of MOS). Speech samples in Exp. I of the ITU-T Supp. 23 database are processed by cascades of various kinds of codecs including G.729, G.726, G.728, GSM-FR, IS-54, and JDC-HR. The correlation coefficients between our PB_MOS and the subjective MOS are given in Table 1. The results show that our algorithm has a comparable performance as the P.563 in estimating MOS of the codec-distorted speech. This work is now further extended to account for frame drop distortions as in Exp. III of the ITU-T Supp. 23 database.

5. CONCLUSIONS

An objective speech quality assessment based on three percepts is proposed and shown to perform comparably to P.563 in estimating MOS of the codec-distorted speech. We show pitch distortions are the most significant degradations produced by speech codecs and are well captured by our defined "naturalness" percept, which weights the most in this work in estimating subjective MOS. This work demonstrates high rate (>100 Hz) temporal modulations are highly correlated with quality of codec-distorted speech.

	P.563	**PB_MOS**
Female 1	0.742	0.683
Female 2	0.759	0.780
Male 1	0.769	0.849
Male 2	0.810	0.840
Average	**0.770**	**0.788**

Table 1. Correlation coefficient between estimated MOS and subjective MOS.

6. ACKNOWLEDGEMENT

This research is partially supported by the National Science Council, Taiwan and partially by the Chunghwa Telecom Co., Ltd. with contract R977047.

7. REFERENCES

[1] "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, 2001.

[2] A.W. Rix, J.G. Beerends, D.-S. Kim, P. Kroon and O. Ghitza, "Objective assessment of speech and audio quality-Technology and applications," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 14, no.6, pp.1890-1901, 2006.

[3] "Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications," ITU-T Recommendation P.563, 2004.

[4] T. Chi, P. Ru, and S.A. Shamma, "Multi-resolution spectro-temporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887-906, 2005.

[5] M. Elhilali, T. Chi, and S.A. Shamma, "A spectro-temporal modulation index for assessment of speech intelligibility," *Speech Communication*, vol. 41, no. 2-3, pp.331-348, 2003.

[6] T. Chi, Y. Gao, C.G. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp.2719-2732, 1999.

[7] "ITU-T coded-speech database," 1998, Supp. 23 to P series rec., ITU-T.

[8] R. Drullman, J. Festen, and R. Rlomp, "Effect of envelope smearing on speech reception," *J. Acoust. Soc. Am.* vol. 95, pp.1053–1064, 1994.

[9] D.-S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no.5, pp.821-831, 2005.

[10] W.M. Parrish, "The concept of naturalness," *Quarterly Journal of Speech*, vol. 37, pp.448-450, 1951.

[11] W. Sanders, C. Gramlich and A. Levine, "The sensitivity of LPC synthesized speech quality to the imposition of artificial pitch, duration, loudness, and spectral contours," *J. Acoust. Soc. Am.*, vol. 64, no. S1, pp.S159, 1978.

[12] A. Ratcliff, S. Coughlin, and M. Lehman, "Factors influencing ratings of speech naturalness in augmentative and alternative communication," *Augmentative and Alternative Communication*, vol. 18, no.1, pp.11-19, Mar. 2002.

計畫成果自評

此研究內容與原計畫書不盡相符。原計畫書是要發展 model-based 的非侵入式語音品質量測(建立乾淨語音之統計模型，將骯髒語音與預估之乾淨語音做比較)，但從去年的研究中給了我們新的想法，似乎我們的聽覺感知運算模型可以試著評估一些語音品質的感知特性，例如理解度、清晰度及自然度，進而發展 parameter-based 的非侵入式語音品質量測。這種 parameter-based 的語音品質量測的想法更直觀且更自然。此想法成功地用來預測各種編碼器所造成的語音品質失真狀況，而其研究成果已發表於重要之國際學術研討會：International Conference on Acoustics, Speech, and Signal Processing (ICASSP) 2009, pp. 4521-4524. 目前我們正延伸此想法，希望能預測一般噪音狀況下之語音品質，而其成果將投稿於國際學術期刊。

2009 ARO (Association for Research in Otolaryngology) MidWinter Research Meeting

冀泰石 (Tai-Shih Chi)
國立交通大學電信工程系

（一）主要任務

The 2009 ARO Mid-Winter Research Meeting was held from Feb. 14 to Feb. 19 at Baltimore, Maryland, USA. There were more than thousands of researchers around the world attending this annual meeting and demonstrating about one thousand of lectures and posters. This ARO annual meeting covers the research areas of biophysics, bio-chemistry, psycho-physics, psycho-acoustics and neuro-physiology of the auditory system of the normal hearing or the hearing impaired people. This meeting provides a good chance for me to have contacts with famous international researchers, learn their newest researches and developments, exchange ideas and explore possibilities for future international collaborations. The outcomes of me attending this research meeting will definitely provide insights in my current or future researches, especially in the development of the next generation hearing aids, which is one of my current projects collaborating with other EE professors in Taiwan.

（二）學術交流及與會經過

(1) The most important gain from this meeting is to establish contacts with international researchers and familiar with their newest researches. For example, the Symposium “Importance of Temporal vs Spectral Fine Structure for Pitch and Speech” in the second day presented many insightful ideas to my current research on speech quality. Bob Shannon introduced three ranges of temporal envelopes well perceived by cochlear implants in the first introductory lecture of the symposium: (1) 2~20 Hz for speech prosody and rhythm; (2) 20~50 Hz for envelope fluctuations; (3) 50~500 Hz for periodicity pitch. These ranges are consistent with our defined temporal modulations conveying three percepts, intelligibility, clarity and naturalness, for speech quality. Bob Carlyon further indicated that pitch coding in normal hearing people is dominated by resolved harmonics while people with cochlear hearing loss have poor reception of the fine structure, hence, impair the temporal integration and the perceived speech quality. Fan-Gang Zeng echoed Bob Shannon with further definition of the “fine structure” as 500~10000 Hz temporal components. He suggested envelopes (5~50 Hz) convey intelligibility and fine structures convey auditory “object” sensation. These lectures support our ideas of investigating speech quality from different regions of temporal envelopes which convey different sensations.

(2) Besides lectures in the symposium, many posters also drew my attentions and provided insightful ideas to another one of my current researches, speech enhancement for digital hearing aids. The group leaded by Philipos Loizou presented two posters about the strategies used in cochlear implants to enhance speech perception. One used a binary masker per cochlear channel to turn off channels with severe SNR compared to a pre-set value. The other approach was to use multi-microphones combined with an adaptive algorithm to reduce the noise. This approach is similar to the beamforming technique used in Phonak high-end hearing aids. In addition to these technique approaches, I found one research done by Stuart Rosen more interesting. They found preserving pitch track would have similar effect of doubling the number of cochlear implant channels (i.e., doubling the spectral resolution of the implant) in boosting speech intelligibility for tonal languages such as Mandarin. This finding is very important and can be used in the development of the next generation hearing aids for Mandarin listeners. It is worth mentioning that this research was done with his formal Ph.D. student from Taiwan, Yu-Ching Kuo, who is now a faculty member in Taipei Municipal University of Education.

(3) In addition, I found those researches done by Christophe Micheyl and Andrew Oxenham are fascinating. Their group tries to tackle the auditory streaming process, perceptually organizing sound sequences, by psychophysical experiments with simple stimuli. Although their results with simple stimuli only confirm the basic theory of the auditory streaming, experiments with more complicated stimuli, such as speech or speech-like modulated signals, are under development now. They also tried to address the perception of temporal fine structure by varying the instantaneous frequency (IF) fluctuations on frequency discrimination tasks. Their results suggested if frequency discrimination is based on temporal fine structure (phase-locking) information, the underlying mechanism is surprisingly immune to IF fluctuations over a wide range of rates. Their results provide basic understanding of the hearing science and are good for modeling the process of human auditory streaming (auditory scene analysis), which is our research proposal under review.

（三）國際交流

During this meeting, I had discussions with Stuart Rosen, from UK, Christophe Micheyl, from US, David Eddins, from US, and Christian Lorenzi, from France. Besides, I had a two-hour meeting with Daniel Pressnitzer, from France, Shihab Shamma and Mounya Elhilali, both from US. We discussed our previous collaborations on timbre analysis and set the goal to submit our results to *Journal of Acoustical Society of America* later this year. Daniel also showed us his new psychophysical results in the reaction times of human in identifying natural sounds. His results showed that temporal cues are quickly adopted by listeners to distinguish among instruments, such as percussion (marimba and vibraphone) from bowed strings (violin and cello). It suggests spectro-temporal patterns may not be used in full to perform fast identification within instruments, only temporal cues (attack time, etc.) are needed for fast identification. It implies people utilize separable temporal and spectral mechanisms for timbre analysis **in general** or such analytical order just stems from years of training from childhood? In our previous collaboration in timbre analysis, we utilized full spectro-temporal features to distinguish instruments.

（四）建議

I have attended this annual meeting for several years. This year, I met some researchers/teachers/doctors from Taiwan, such as Yu-Ching Kuo from Taipei Municipal University of Education. I found that more and more hearing researches are proposed for tonal languages, such as Mandarin, perhaps due to the economic strength of China with its massive population. In this annual meeting, I saw more and more students from China working as Ph.D. students in US participating this meeting. We have to start the full spectrum of hearing researches in Taiwan quickly while we still have the advantages, high tech plus native language. Time is gaining on us!