Issue 9

April 2015

APSIPA NEWSLETTER

Asia-Pacific Signal and Information Processing Association Newsletter

In This Issue

We introduce you APSIPA newsletter issue-9 with many interesting articles and reports in addition to the regular columns. We first present two reports from the technical committees of Wireless Communications and Networking (WCN) and Signal and Information Processing Theory and Methods (SIPTM). We intend to introduce the technical committee reports from the other technical committees in the coming APSIPA newsletter issues. We are aiming toward promoting APSIPA technical committees to attract more members as well as help channeling paper submissions to APSIPA conferences and Journal.

The articles presented in APSIPA newsletter are to prosper knowledge in important research areas in a quick and condensed mode. There are four articles in this issue sequenced as follow:

- Matlab real-time tools for speech and signal processing education
- Affective Analysis of Music Signals using Acoustic Emotion Gaussians: A Brief Overview
- Post-Filter Using Modulation Spectrum as a Metric to Quantify Over-Smoothing Effects in Statistical Parametric Speech Synthesis
- Interference Suppression Schemes for Radio over Fiber Simultaneously Transmitted with 10 Gbps On-Off Keying

We also have in this issue an important announcement about the release of the first publicly released Mandarin-English code-switching speech database for automatic speech recognition research which is available through LDC.

We hope you find this issue useful and please feel free to send us comments and suggestions to improve our newsletter to benefit as much as possible our APSIPA members.

Waleed Abdulla
APSIPA Newsletter Editor-in-Chief

WCN Technical Committee Report	Page 2
SIPTM Technical Committee Report	Page 3
MATLAB Real-time Tools for Speech	Page 5
Affective Analysis of Music	Page 11
Post-filter Using Modulation Spectrum	Page 14
Interference Suppression Schemes	Page 17
SEAME Speech Database	Page 21

Wireless Communications and Networking (WCN) Technical Committee Report

Tomoaki Ohtsuki

It is our great pleasure to introduce Wireless Communications and Networking (WCN) Technical Committee (TC). Current officers are TC Chair: Tomoaki Ohtsuki, Keio University (Japan), Vice-Chair: Sumei Sun, I2R (Singapore), and Secretary: Takahiko Saba, Chiba Institute of Technology.

WCN-TC sponsors papers, participates in organization of conferences, and promotes technical workshops on those aspects of communications that pertain to the innovation, development and application of algorithms and electronic and photonic devices or subsystems for generation, processing, storage, transmission, recovery, and presentation of communications signals. In addition, the committee has professional development goals for committee members and other practitioners working in the above areas.

1. Fields of Interest

The fields of interest of the WCN TC shall be, but not limited to, the following:

- Signal processing in wireless communications, including physical layer, medium access, and networking functions
- Signal processing architecture and circuits for wireless communications and networks
- Signal processing for applications of wireless mobile communications and networks
- Signal representation, theory, and processing toward innovative wireless communications and networks, and their applications

2. TC Membership

- The TC members shall be researchers in the fields of interest with good standing, and be elected by current members of the TC.
- The WCN TC shall consist of maximum 50 members plus the Chair, the Vice Chair, the Secretary, and the Past-Chair. All TC members are voting members. Each member shall ordinarily serve two consecutive three-year terms, but membership may be terminated at any time if the Chair agrees that the member has not fulfilled his/her TC responsibilities. A member whose two three-year terms expire will be eligible for re-election only after a one-year pause.
- Each member of the TC is expected to participate in at least 75% of the votes. Members not meeting their obligations will not be considered for renewal of their first three-year term. Renewal of the first three-year term is subject to a vote by the TC members, and is conducted along with the election of new members as outlined in WCN Bylaw 5, that is available on APSIPA WCN website.
- Memberships expire at the end of the third year.

3. Member Election Procedure:

- The TC Chair and Vice-Chair are responsible to publicize the rules of nomination and prepare a ballot for election.
- Nominations are open to all researchers in the field of interest with good standing. The nominator is responsible for contacting the nominee before the nomination and provides a succinct case for the candidate.
- The candidates who receive the highest number of votes will be elected. In the event of a tie, a swift election is conducted. The quorum for a winning election is 50% of the TC membership.
- The election will be conducted via e-mail, unless another means is proposed and approved by the TC.

4. Supporting Activities

As appropriate, WCN-TC will be active in all of APSIPA's activities. This will include APSIPA hosting conferences (such as APSIPA Annual Summit and Conference) by providing representatives from their respective Technical Program Committees, by soliciting assistance from its membership to provide professional review of submitted papers, and by organizing mini-conferences, symposia, panels, short courses, tutorials, etc., as deemed appropriate by the Society and WCN-TC. Further, WCN-TC can individually organize workshops and conferences. Also due to the broad nature of committee activities, collaborative sessions with other committees will be sponsored and heartily encouraged.

WCN-TC will further support APSIPA Transaction and standards activities by soliciting volunteers as authors and editors, submitting proposals, and identifying committee members from its membership. WCN-TC will seek ways to increase active participation of its members in information exchange related to the charter of this Committee, such as: stimulating Feature Topics and Special Issues of APSIPA Transaction; and by sponsoring workshops, tutorials, short courses, panel sessions, etc. on special topics.

APSIPA Signal and Information Processing Theory and Methods Technical Committee Report

Anthony Kuh and Akira Hirabayashi

This is a note of activities by the Signal and Information Processing Theory and Methods (SIPTM) Technical Committee (TC). The SIPTM TC is an active TC promoting the rich area of signal and information processing theory and methods. This area has grown in recent years with much research and education in the areas of theory, algorithms, and applications. Areas range from traditional topics such as digital filters, statistical signal processing and adaptive signal processing to newer topics in applying signal and information processing to big data and smart grids. The SIPTM TC promotes activities through organization of special and regular sessions at the APSIPA Annual Summit and Conference (ASC), through publication of journal articles in the APSIPA Transactions on Signal and Information Processing, through articles such as this one in the APSIPA

Newsletter discussing activities, and encouraging researchers and educators in signal and information processing to join our community.

The last APSIPA ASC was held in Siem Reap, Cambodia. We held a SIPTM TC meeting on December 10th. The meeting was attended by Oscar Au, Mrityunjoy Chakraborty, Akira Hirabayashi, Anthony Kuh, and Yili Xia. At the meeting we discussed recent activities of the SIPTM TC and also ways we could contribute to APSIPA including publicizing SIPTM TC activities. There was also a succession of leadership with Akira Hirabayashi succeeding Anthony Kuh as the new chair of the SIPTM TC. SIPTM had several special and regular sessions at the 2014 APSIPA ASC including special sessions on ``Signal and Information Processing Applications of Smart Grids and Energy'' and ``Machine Learning Algorithms and Applications in Signal Processing''.

In the APSIPA Transactions on Signal and Information Processing we currently have a special ongoing issue on ``Signal and Information Processing for the Smart Grid''. In 2014 three articles from this special issue appeared and there are currently more papers in the review process. We look forwards to more contributions from the SIPTM community for both special issues and journal submissions to the APSIPA Transactions on Signal and Information Processing.

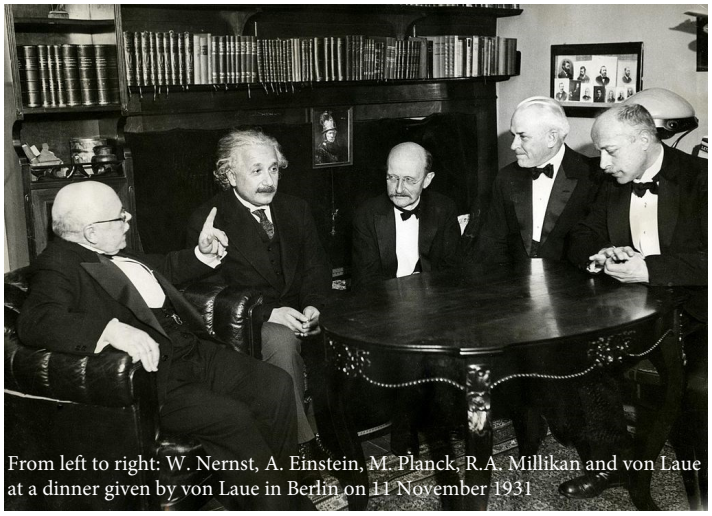
We welcome researchers and educators in signal and information processing to join our TC by contacting either Akira at akirahrb@media.ritsumei.ac.jp or Anthony at kuh@hawaii.edu.

Science Quotes

Science cannot solve the ultimate mystery of nature. And that is because, in the last analysis, we ourselves are part of nature and therefore part of the mystery that we are trying to solve.

— **Max Planck** (Born 23 Apr 1858; died 4 Oct 1947)
From: Where is Science Going, Max Planck (1933).

There is a story that once, not long after he came to Berlin, Planck forgot which room had been assigned to him for a lecture and stopped at the entrance office of the university to find out. Please tell me, he asked the elderly man in charge, 'In which room does Professor Planck lecture today?' The old man patted him on the shoulder 'Don't go there, young fellow,' he said 'You are much too young to understand the lectures of our learned Professor Planck'.


From left to right: W. Nernst, A. Einstein, M. Planck, R.A. Millikan and von Laue at a dinner given by von Laue in Berlin on 11 November 1931

From: Men Who Made a New Physics: Physicists and the Quantum Theory,
Barbara Lovett Cline (1987).

Matlab realtime tools for speech and signal processing education

Hideki Kawahara*
*Faculty of Systems Engineering, Wakayama University, Wakayama, Wakayama, Japan
E-mail: kawahara@sys.wakayama-u.ac.jp Tel: +73-457-8461

***Abstract*—Interactive realtime tools for education were developed making use of Matlab GUI functions. They are a) a realtime spectrogram with narrowband, wideband and perceptual frequency resolution axes, b) a realtime spectrum monitoring followed by an interactive spectrogram display with sound playback of a rectangular time-frequency region and scrubbing inspection, c) realtime vocal tract shape display of input sounds, and d) realtime F0 extraction with event detection and display. In addition to these tools, animation generator functions are prepared for understanding fundamentals of Fourier transform and digital signal processing.**

I. Introduction

Extraordinary advancement in computational power of personal machines makes it possible to use these machines as interactive learning tools for understanding not-very-intuitive mathematical concepts underlying speech, hearing and digital signal processing. In addition to this computational power, introduction of modern high-level software development systems greatly reduces necessary time an d effort to implement such tools. This article introduces tools we have developed for training undergraduate and graduate students. The tools, with their Matlab source codes, are freely available from the author's web page [1].

II. Before introducing realtime tools

Students have to have some acquaintance to the concepts such as time, frequency and waveform before introducing tools described in this article. In the very beginning of the introduction lecture of media technologies for the first year undergraduate students, I prepared animation materials shown in Fig. 1. The top left movie introduces the relation between complex exponential function and sinusoids, Euler's equation. The top right movie introduces Fourier transform of periodic signals as a Ptolemaic system by morphing settings of harmonic amplitude and phase shown in the lower part of the movie. In addition to these movies, a movie with sound, shown in the bottom left panel, is presented to make students acquire/experience relation between pitch, timbre, harmonic structure and waveform. For graduate students, I prepared one additional movie, shown in the bottom right panel, for introducing relative phase of harmonic signals on timbre perception and waveform, using a cyclic transition from cosine phase, sine phase, alternating phase [2], and Schroeder phase [3]. This snap shot shows Schroeder phase.

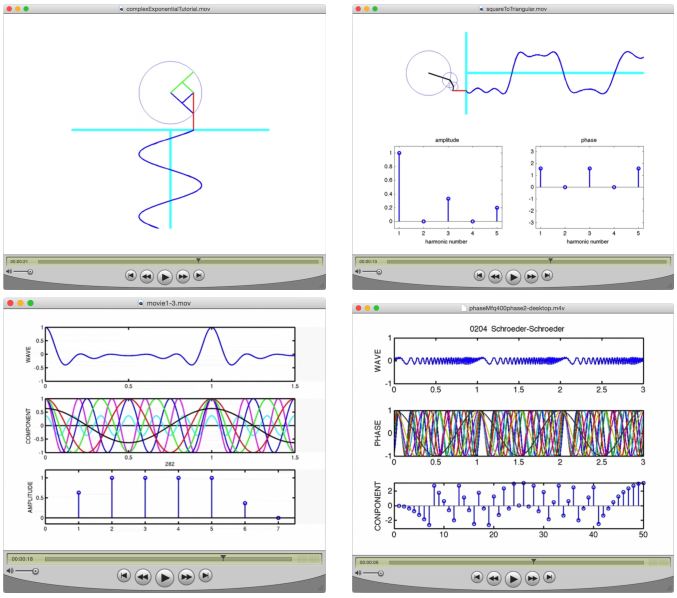


Fig. 1. Introduction movies to (top left) complex exponent, (top right) Fourier series, (bottom left) harmonic structure and timber, and (bottom right) relative phase on timber and waveform.

All these movies were made using Matlab [4] scripts and QuicTime Player 7 pro. The moves shown in Fig. 1 and the Matlab scripts archive are accessible from my site. But please note that this cumbersome procedure for making movies with sounds is already outdated today. It can be replaced by Matlab GUI-based programming. The following sections also show such examples.

III. Realtime Matlab tools

This section introduces each Matlab realtime tool with examples of basic operation and function. It also provides practical hints for implementing such tools using Matlab GUI functions. These tools only require the signal processing toolbox and do not use platform dependent codes.

A. Realtime spectrum analysis and recording tool

This tool has two modes of operation; realtime monitoring and recording mode and interactive inspection mode.

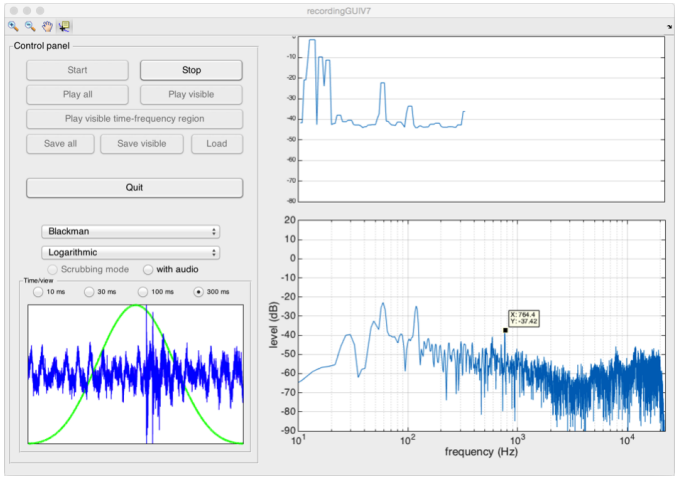


Fig. 2. Screen capture of the realtime spectrum analyser mode when monitoring the background noise in my office. This illustrates realtime monitoring of the waveform segment, the time windowing function and the calculated power spectrum using log-frequency axis.

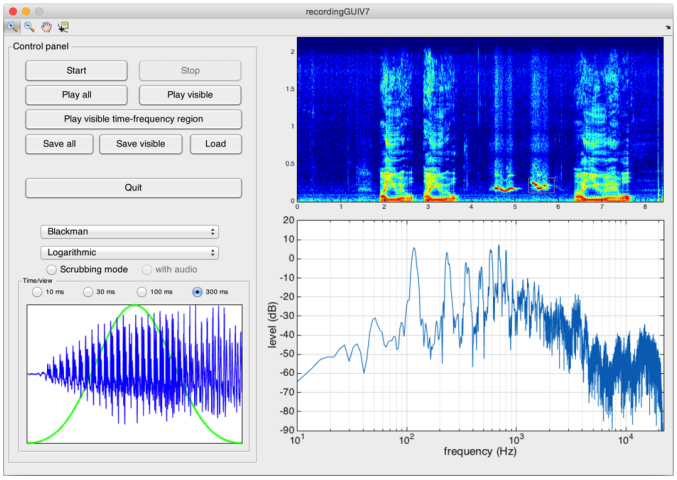


Fig. 3. Initial display of interactive inspection mode. Three segments of Japanese vowel sequences and two whistling sounds were acquired. Note that the zoom-in tool is highlighted in the tool bar. The second whistling region is being selected using rubberband selector around 5.6 s. The rubberband is made visible by magnifying the region of this spectrogram.

1) Realtime monitoring and recording mode: Figure 2 shows a screen capture of this tool in its realtime monitoring and recording mode. The bottom left panel updates waveform in realtime and the bottom right panel also updates the power spectrum of the waveform segment shaped by the time windowing function that is also displayed in the bottom left panel. In this mode, top right panel displays the peak level of each analysis segment. The clipping level is set 0 dB in this mode.

This GUI uses default toolbar icons shown in the top left corner of the tool. They are zoom-in, zoom-out, panning and data cursor. In this snapshot, a data cursor is attached to one of a salient peak of the power spectrum plot and updating the attached coordinate values in realtime. The peak was caused by the noise from the desiccator for measuring microphone storage. These icon tools can be toggled by clicking while monitoring realtime.

The available windowing functions are Blackman, Hamming, Hanning, Bartlett, Rectangular [5] and Nuttall [6] windows. The preset window lengths are 10, 30, 100 and 300 ms. The frequency axis can be selected from the following list; Logarithmic, Linear full scale, Linear 1000 Hz, Linear 8000 Hz and Linear 4000 Hz. This figure shows Linear full scale. These settings can also be changed on the fly.

In this mode, only Stop and Quit buttons are enabled. Clicking Stop button terminates data acquisition and displays the spectrogram of the acquired data. Displaying spectrogram initiates interactive inspection mode.

2) Interactive inspection mode: Figure 3 shows a snap shot of inspection mode. Top right panel shows the spectrogram of the all acquired data in the time domain and 0 Hz to $f_s/2$ in the frequency domain, where f_s represents the sampling frequency. The analysis window is a 15 ms Blackman window with 2 ms frame update. The pseudo colour dynamic range is 80 dB from the peak power spectrum level. In this figure the second whistling region around 5.6 s is being selected using a rubberband selector. Figure 4 also shows the zoomed region (This display show scrubbing mode. This is explained later.).

The GUI toolbar icons are also usable in this mode. Spectrogram, waveform and power spectrum display panels can be zoomed, panned and by using the data cursor coordinate values can be monitored.

In this mode, buttons other than “Stop” is enabled. The “Play all” button plays back all data acquired in the input buffer. The “Play visible” button plays back the data, which is displayed in the zoomed spectrogram. The “Play visible time-frequency region” plays back the sound which corresponds to the selected time-frequency region by zooming operation. This operation is implemented using FIR filter design based on the visible frequency axis information of the spectrogram display panel. The “Save all” button saves whole acquired data to a wave file. The “Save visible” button saves only the visible portion of the data. The “Load” button replaces the whole acquired data with the content of the file selected by Matlab GUI’s file input dialogue.

This “Play visible time-frequency region” function is useful for demonstrating timbre differences between different frequency regions and age related hearing loss. It can be instructive to playback so-called “mosquito sound” using this function. The majority of the sound is located around 16 kHz. Zooming the region around 16 kHz sounds like silence for (most of the) teachers but not for students. It sounds irritating tweets for teenagers, majority of the first year undergraduate students.

3) Scrubbing of spectrogram display: When the “Scrubbing mode” radio button is on, the spectrum slice, the waveform segment and the corresponding sound of a specific portion of the spectrogram can be inspected by scrubbing the spectrogram using the scrubbing cursor. As mentioned before, the

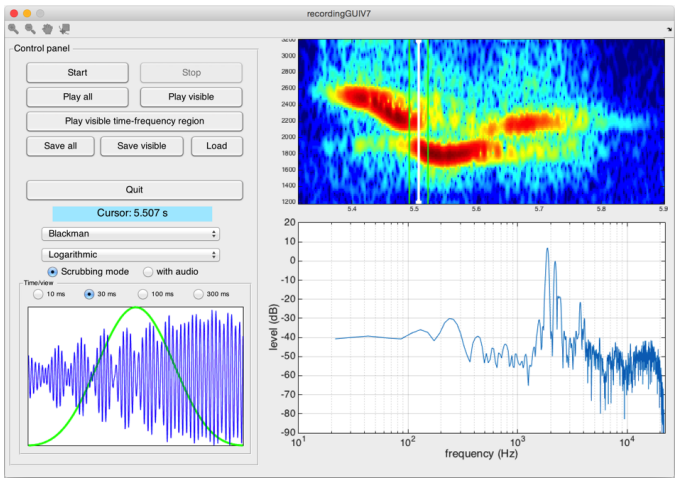


Fig. 4. Zoomed spectrogram in scrubbing mode ON. The zoomed region shows a transitional mode of whistling, where two oscillation modes are co-existing and produce beating in the waveform.

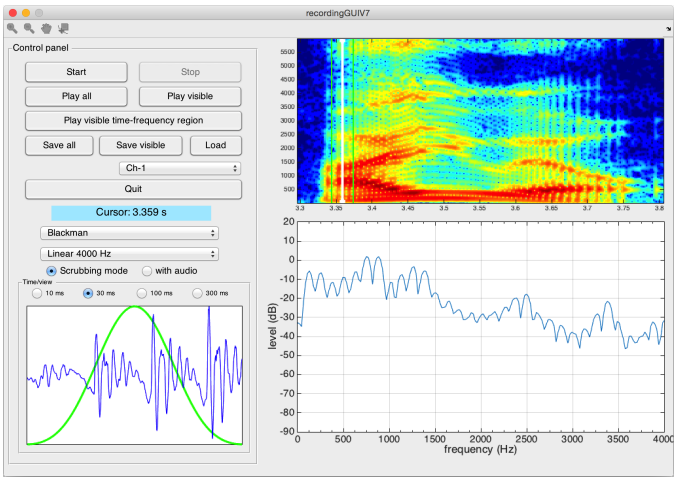


Fig. 5. Zoomed spectrogram in scrubbing button ON. The scrubbing cursor is located at the initial part of a Japanese vowel sequence /aiueo/ spoken by a male speaker. The power spectrum display uses the linear frequency axis from 0 Hz to 4000 Hz.

zoomed whistling display in Fig. 4 shows a snap shot in this mode.

The user can drag the scrubbing cursor (white thick vertical line in the spectrogram) to view any part of the displayed spectrogram. Two green lines flanking the scrubbing cursor indicates segment boundaries currently analysed. The waveform panel updates the display synchronised with the drag movement. The power spectrum display also is updated. Similar to the realtime monitoring mode, the window shape and the length, and frequency axis can also be changed on the fly. (Figure 5 shows a snap shot with linear frequency axis up to 4000 Hz with 30 ms analysis window length. This setting is relevant for demonstrating vowel and formant relations.) Scrubbing radio button is enabled to be selected only when all GUI toolbar icons are deselected. By setting "with audio" radio button ON, sound can also be played back while scrubbing. Please use this "with audio" function carefully in classroom situation, because using this mode for speech makes it sound like stuttering. One of students reported after my lecture that it hurt him, a stutterer.

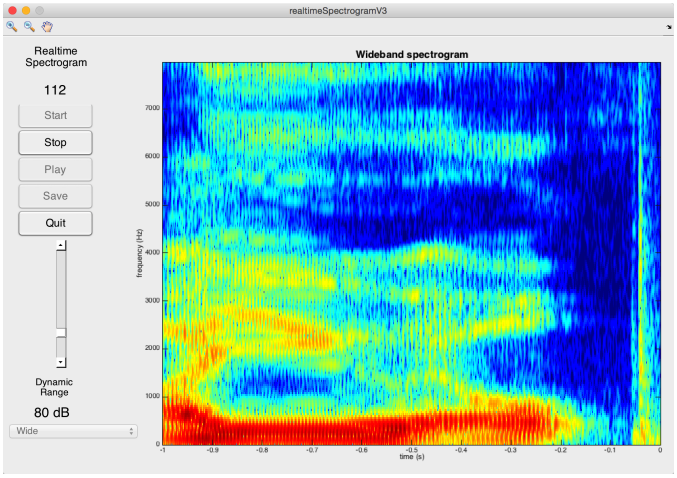


Fig. 6. Realtime spectrogram in "wide band" display mode. The spectrogram is continuously scrolling from right to left. A Japanese vowel sequence /aiueo/ spoken by a male and a click sound are shown.

B. Realtime spectrogram tool

Figure 6 shows a snap shot of the realtime spectrogram tool in its "wide" band spectrogram display mode. The duration of the displayed spectrogram is one second. The spectrogram is continuously scrolling from right to left. The windowing function is Nuttall window with 10 ms in length and 0.5 ms frame update interval. The display update interval is 100 ms in this implementation. In "narrow" band spectrogram display mode, the window length and the frame update interval are set 80 ms and 7 ms respectively. The duration of the displayed spectrogram is increased to two seconds. In wide band spectrogram, the frequency rage is set from 0 Hz to 8000 Hz. In narrow band spectrogram, the frequency rage is set from 0 Hz to 4000 Hz. These modes are designed for introducing basics of acoustics phonetics and the wide band display is used mainly for showing formant trajectories and the narrow band display is used mainly showing prosody, in other words, F0 (fundamental frequency) trajectories and harmonic structure.

This tool also has several types of simulated filter bank mode. They are a) ERB_N number [7] frequency axis with 1 ERB_N filter bank, b) Bark [8] filter bank and c) one-third octave filter bank. Figure 7 shows the ERB filter bank. Please note that these simulated filter bank analyses are calculated from the narrow band spectrogram and the temporal resolution of these displays are highly smeared in the higher frequency region. Also the group delay caused by filtering is not represented. The latter problem is severe in the lower frequency region. Even with these limitations, simulated filter bank mode is useful for demonstrating differences between engineering representations, such as wide band, narrow band

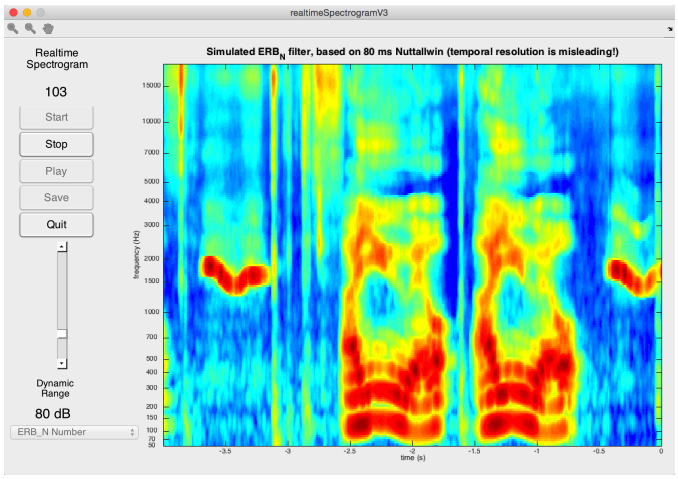


Fig. 7. Realtime spectrogram in "ERB_N number" display in simulated filter bank analysis mode. The spectrogram is continuously scrolling from right to left. Two segments of Japanese vowel sequence /aiueo/ spoken by a male and two segments of whistling are shown.

and one third octave, and perceptual representations, such as ERB and Bark. Please note that ERB_N number closely matches with the frequency allocation on the basilar membrane in the cochlea of human [9].

Display modes selection menu in the bottom left corner is enabled when "Stop" button is clicked to suspend running spectrogram mode. After selection of the display mode, running spectrogram starts. The vertical slider is used to sed the dynamic range of pseudo colour display. The gain of the automatic gain controller is reset to prevent clipping of the signal and linearly slowly increases while no clipping happens.

C. Event display tool

Figure 8 shows snap shots of the event display tool. This tool is mainly designed to serve as a template of realtime programming using Matlab GUI functions. This tool calculates one third octave band levels, modified correlation based F0 and maximum correlation (a variant of frequency domain methods [10], dividing power spectra by their frequency-smoothed versions followed by base-band weighting and inverse Fourier transform to yield modified autocorrelations, 1970s' technology), segmental power, and segmental kurtosis. It also detects acoustic event based on the maximum correlation, the power jump and the kurtosis value. These are calculated simultaneously triggered by timer interruption (interval is 50 ms in this tool) associated with the audio recording object in Matlab.

The central plot of each panel displays F0 values as green dots. The frequency axis is represented in either the musical note scale or the logarithmic frequency. These display scales can be changed on the fly by clicking the radio button shown in the left side of the GUI.

The top plots show the dB power, acoustic events detected by the kurtosis, events detected by the onset of segmental power, and periodic events such as voiced sounds. The three bars in the bottom left represents the power, the kurtosis and

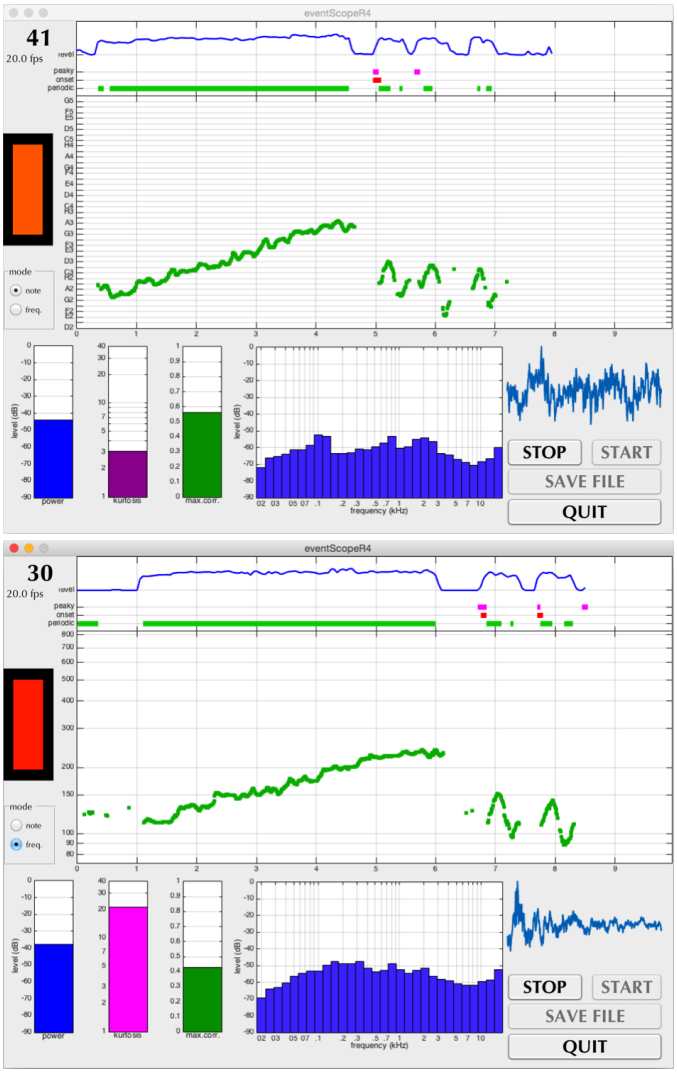


Fig. 8. Realtime event detection tool. The initial part shows a musical scale sang in vowel /a/ sound and the latter part shows segments of vowel sequence /aiueo/ by a male. The top panel shows F0 trajectories using musical note axis. The bottom panel shows them using the logarithmic frequency axis. The frequency axes mode can be changed on the fly.

the maximum correlation of each analysis frame. As shown in the bottom panel of Fig. 8, increase in the colour saturation level indicate event detection by the corresponding bar(s). In this example, the event is detected by the high kurtosis level.

The one third octave band level display and the waveform display are updated in the same interval, every 50 ms. The face colour of the left rectangular box is determined by mapping relative power levels of lower, middle and higher frequency bands to RGB colour levels.

D. Vocal tract shape tool

Figure 9 shows the realtime display of the vocal tract shape of the input signal. The top right plot shows the logarithmic area function. The horizontal axis represents the distance from the lip opening. The three dimensional vocal tract shape in the bottom right plot is drawn using a constant volume

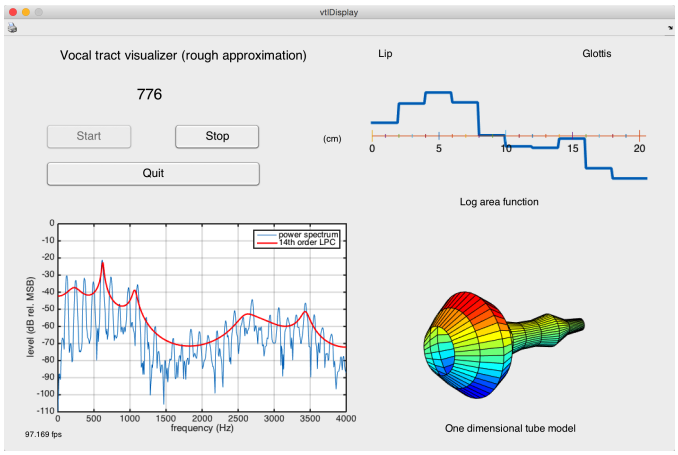


Fig. 9. Realtime display of the vocal tract area function.

constraint on the vocal tract cavity. The vocal tract area function is derived from LPC [11], [12] coefficients through reflection coefficients [13], [14], in other words PARCOR coefficients [15]. A pre-processing stage consisting of high frequency emphasis (differentiation) and equalisation using first order FFT cepstrum coefficient is used for compensating effects of glottal source spectral shape and the radiation characteristics from lip opening to the open space.

The three dimensional view can be rotated in three dimensional space on the fly in any mode. This rotation functionality is default for Matlab graphics objects. The bottom left plot shows the short term power spectrum and the spectral envelope based on LPC analysis results. The window type, the length and the frame shift are Blackman, 30 ms and 8 ms respectively. The sampling frequency was set 8000 Hz. The LPC analysis results using the analysis order 14 is used to display the spectrum envelope. For vocal tract shape analysis, the analysis order 10 is used, since needs of extra analysis order is removed by the pre-processing stage.

IV. Implementation

The open GUI layout editor of Matlab [4], "guide" was used in GUI design and programming of these tools. Realtime update of displays are written as event handlers which are triggered by the timer interruption associated with the Matlab audio input object. Interactive inspections such as scrubbing are also written as event handlers. Button up, button down, motion handlers are dynamically assigned or released depending on events for enabling interactive scrubbing and synchronised update of displayed plots. These handlers are assigned to some of graphic objects such as the scrubbing cursor.

V. Discussion

The audio output object of Matlab also can define timer interruption and "guide" provides means to write event handlers for it. Using this timer interruption, animation movies with sound output introduced in Fig. 1 can easily be implemented as interactive realtime tools. It is generally painful to inspect

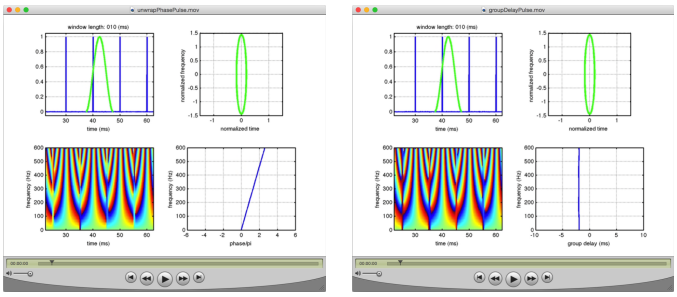


Fig. 10. Animation movies for introducing group delay. (Left plot) unwrapped phase of a windowed pulse train. (Right plot) group delay of a windowed pulse train.

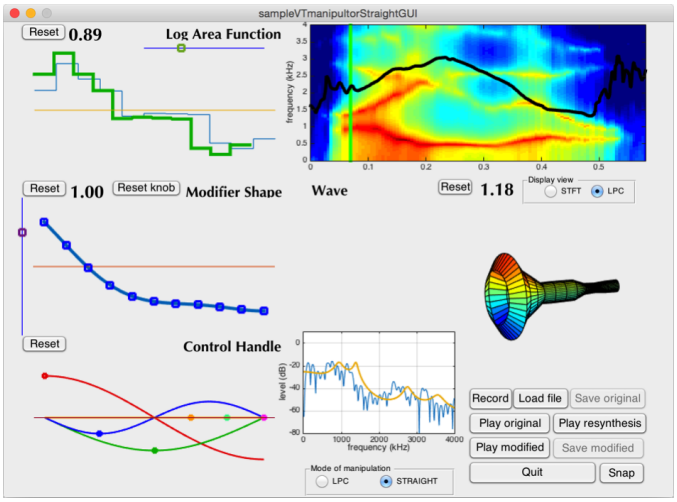


Fig. 11. Vocal tract-based STRAIGHT parameter manipulation GUI.

30 sheets of graph plot and to find underlying phenomena but it is easy or sometimes fun to observe a scientific visualisation movie of one second long with 30 frames per second update to find what kind of change is happening. In this respect, interactive realtime animation tools are promising for assisting speech and signal processing education.

Figure 10 shows one such example introducing the concept of group delay. In both panels, the top left plot shows the pulse train and windowing function. The bottom left plot shows the phase spectrogram. The horizontal axis of each plot is aligned. In the left panel, the bottom right plot shows the unwapped phase of the calculated short term Fourier transform. The same plot in the right panel shows the group delay. Comparing these two movies help students acquire an physically meaningful interpretation of group delay as a centroid of energy in each frequency.

The Matlab source codes provide templates for designing tools for serious research. Figure 11 shows one such example. This GUI enables parameter modification of STRAIGHT, speech analysis, modification and resynthesis system [16], [17], based on an intuitive representation, vocal tract shape [18]. This GUI is implemented in two days by modifying the code of the vocal tract shape tool. The top

right spectrogram display, the bottom center power spectrum and LPC spectral envelope display, the top left logarithmic area function display, and three dimensional vocal tract shape display are updated synchronously to manipulation of vocal tract shape by using tools in the left side of the GUI. Details of this tool will be presented elsewhere.

VI. Conclusions

A set of realtime tools implemented using GUI functions of Matlab is introduced. Also a set of introduction movies for media technologies is introduced. Their Matlab source codes are freely accessible from the author's web page [1].

Acknowledgment

The author appreciates to the APSIPA newsletter editor Prof. Abdulla for suggesting this submission for sharing codes and practice in signal processing education. This article serves as a subtext of one of the topics in my lecture as a 2015-2016 Distinguished Lecturer of APSIPA.

References

[1] H. Kawahara. Hideki kawahara, professor wakayama university. [Online]. Available: http://www.wakayama-u.ac.jp/˜kawahara/index_e.html

[2] R. D. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Am.*, vol. 82, no. 5, pp. 1560–1586, 1987.

[3] M. R. Schroeder, "Synthesis of low-peak-factor signals and binary sequences with low autocorrelation," *IEEE Trans. Information Theory*, vol. 16, no. 1, pp. 85–89, 1970.

[4] Matlab, *8.4.0.150421 (R2014b)*. Natick, Massachusetts, USA: The MathWorks Inc., 2013.

[5] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE*, vol. 66, no. 1, pp. 51–83, 1978.

[6] A. H. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Trans. Audio Speech and Signal Processing*, vol. 29, no. 1, pp. 84–91, 1981.

[7] B. C. Moore and B. R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[8] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *The Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1523–1525, 1980.

[9] D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *The Journal of the Acoustical Society of America*, vol. 87, no. 6, pp. 2592–2605, 1990.

[10] W. Hess, *Pitch Determination of Speech Signals: Algorithms and Devices*. Springer-Verlag, 1983.

[11] F. Itakura and S. Saito, "A statistical method for estimation of speech spectral density and formant frequencie," *Electro. Comm. Japan*, vol. 53-A, no. 1, pp. 36–43, 1970, [in Japanese].

[12] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *J. Acoust. Soc. Am.*, vol. 50, no. 2B, pp. 637–655, 1971.

[13] H. Wakita, "Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 3, pp. 281 – 285, jun 1979.

[14] J. Schroeter and M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 133–150, Jan 1994.

[15] F. Itakura and S. Saito, "On the optimum quantization of feature parameters in the parcor speech synthesizer," in *Conference Record, 1972 International Conference on Speech Communication and Processing, Boston, MA*, 1972, pp. 434–437.

[16] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[17] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," in *Proc. ICASSP 2008*, 2008, pp. 3933–3936.

[18] A. Arakawa, Y. Uchimura, H. Banno, F. Itakura, and H. Kawahara, "High quality voice manipulation method based on the vocal tract area function obtained from sub-band LSP of STRAIGHT spectrum," in *Proc. ICASSP 2010*, March 2010, pp. 4834–4837.

Affective Analysis of Music Signals using Acoustic Emotion Gaussians: A Brief Overview

Ju-Chiang Wang,⋆ Yi-Hsuan Yang,⋆† and Hsin-Min Wang⋆†
⋆Institute of Information Science, Academia Sinica, Taipei, Taiwan
†Research Center for IT Innovation, Academia Sinica, Taipei, Taiwan
Emails: {asriver, yang, whm}@iis.sinica.edu.tw

Introduction

We provide in this letter a brief overview of a recently proposed generative model called Acoustic Emotion Gaussians (AEG) for emotion-based music signal processing and information retrieval [3]–[8]. The idea is to present the possible affective responses to a music clip by a *probability* distribution, so as to account for the subjective nature of emotion perception. The term *affective response* in this letter refers to the emotion people perceive as being expressed in a music clip. Moreover, we describe emotions in terms of *valence* (or pleasantness; positive or negative affective states) and *arousal* (or activation; energy level), the two most important dimensions of emotion [2], [9], [10]. For example, happiness is an emotion state associated with a positive valence and a high arousal, while sadness is an emotion state associated with a negative valence and a low arousal. The valence-arousal (VA) space is viewed as an Euclidean space and any point in the VA space can be considered as a specific emotion state.

The name of the AEG model comes from its use of multiple Gaussian distributions to model the affective content of music. The algorithmic part of AEG has been first introduced in [6], along with the preliminary evaluation of AEG for emotion-based music annotation (aka music emotion recognition, or MER). More details about the analysis part of the model learning of AEG can be found in a recent article [8]. Due to the parametric nature of AEG, model adaptation techniques have also been proposed to personalize an AEG model in an online, incremental fashion, rather than learning from scratch [1], [7]. The application of AEG to emotion-based music video generation and emotion-based music retrieval can be found in [4] and [5], respectively. Moreover, as shown in [3], the AEG model can be extended to explore the connection between emotion dimensions and discrete emotion categories for visualization and retrieval purposes.

In what follows, we briefly describe the AEG model, and then introduce its five possible applications. The source codes for implementing AEG can be obtained from the link: http://slam.iis.sinica.edu.tw/demo/AEG/.

The AEG Model

As Figure 1 shows, AEG involves the generative process of VA emotion distributions from audio signals. To start the generative process of AEG, we first learn an *acoustic GMM* as the bases to represent a music clip,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k A_k(\mathbf{x} \mid \mathbf{m}_k, \mathbf{S}_k)\,, \tag{1}$$

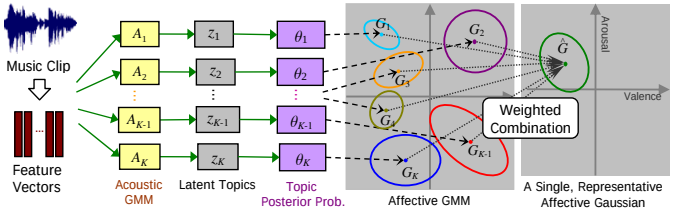


Fig. 1. Illustration of the generative process of the AEG model.

where $A_k(\cdot)$ is the k-th component Gaussian, and π_k, $\mathbf{m}_k$, and $\mathbf{S}_k$ are its corresponding prior weight, mean vector, and covariance matrix, respectively. Suppose that we have an emotion annotated corpus $\mathcal{X}$ consisting of N music clips $\{s_i\}_{i=1}^{N}$. Given a set of short-time feature vectors $\{\mathbf{x}_{i,t}\}_{t=1}^{T_i}$ extracted from a clip s_i, we first compute the posterior probability for each feature vector,

$$p(A_k \mid \mathbf{x}_{i,t}) = \frac{A_k(\mathbf{x}_{i,t} \mid \mathbf{m}_k, \mathbf{S}_k)}{\sum_{h=1}^{K} A_h(\mathbf{x}_{i,t} \mid \mathbf{m}_h, \mathbf{S}_h)}\,. \tag{2}$$

Then, the clip-level topic posterior probability $\theta_{i,k}$ of s_i can be approximated by

$$\theta_{i,k} \leftarrow p(z_k \mid s_i) \approx \frac{1}{T_i} \sum_{t=1}^{T_i} p(A_k \mid \mathbf{x}_{i,t})\,. \tag{3}$$

We put the values of $\{\theta_{i,k}\}_{k=1}^{K}$ in a single vector $\boldsymbol{\theta}_i \in \mathbb{R}^K$ called *topic posterior vector* to represent the audio content.

In view of the subjectivity of emotion perception, a music clip in the labeled dataset is usually annotated by multiple subjects. Let $\mathbf{e}_{i,j} \in \mathbb{R}^2$ (a vector indicating the VA values) denote the annotation of s_i given by the j-th subject, and U_i denotes the number of subjects who have annotated s_i. We assume that each annotation $\mathbf{e}_{i,j}$ can be generated from an *affective GMM* weighted by its topic posterior vector $\boldsymbol{\theta}_i$,

$$p(\mathbf{e}_{i,j} \mid \boldsymbol{\theta}_i) = \sum_{k=1}^{K} \theta_{i,k} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,, \tag{4}$$

where $G_k(\cdot)$ is a bivariate Gaussian, and $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ are the mean vector and covariance matrix of the k-th component (i.e., latent topic). Then, we obtain the log-likelihood function of total annotations over $\mathcal{X}$:

$$L = \log \sum_i \sum_j \sum_k \theta_{i,k} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\,. \tag{5}$$

To learn the parameters of the affective GMM, we adopt the EM algorithm to maximize Eq. 5 with respect to $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$.

In the E-step, we compute the posterior probabilities:

$$p(z_k \mid \mathbf{e}_{i,j}) = \frac{\theta_{i,k} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{h=1}^{K} \theta_{i,h} G_k(\mathbf{e}_{i,j} \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)} . \quad (6)$$

In the M-step, we obtain the updating forms,

$$\boldsymbol{\mu}'_k \leftarrow \frac{\sum_i \sum_j p(z_k \mid \mathbf{e}_{i,j}) \mathbf{e}_{i,j}}{\sum_i \sum_j p(z_k \mid \mathbf{e}_{i,j})} , \quad (7)$$

$$\boldsymbol{\Sigma}'_k \leftarrow \frac{\sum_i \sum_j p(z_k \mid \mathbf{e}_{i,j}) (\mathbf{e}_{i,j} - \boldsymbol{\mu}'_k)(\mathbf{e}_{i,j} - \boldsymbol{\mu}'_k)^T}{\sum_i \sum_j p(z_k \mid \mathbf{e}_{i,j})} . \quad (8)$$

A thorough analysis of the learning process of the AEG model can be found in [8].

Music Emotion Recognition

AEG predicts the emotion distribution of an unseen clip by using its topic posterior $\hat{\boldsymbol{\theta}} = \{\hat{\theta}_k\}_{k=1}^{K}$ on the learned affective GMM. We can also use a single, representative Gaussian $G(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ to approximate the weighted affective GMM [6]:

$$\hat{\boldsymbol{\mu}} = \sum\nolimits_k \hat{\theta}_k \boldsymbol{\mu}_k , \quad (9)$$

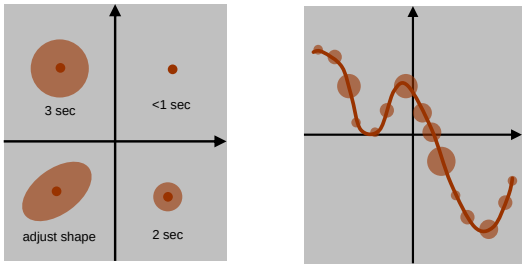
$$\hat{\boldsymbol{\Sigma}} = \sum\nolimits_k \hat{\theta}_k \left(\boldsymbol{\Sigma}_k + (\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}) (\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}})^T \right) . \quad (10)$$

The accuracy of MER can be measured by computing the KL divergence between the groundtruth (i.e., human labeled) and the predicted ones, or by comparing the mean VA values. Evaluations on two emotion-labeled datasets have validated the effectiveness of AEG over prior arts for MER [8].

Personalized Music Emotion Recognition

Personalization is important for practical emotion-based music applications. As AEG is a probabilistic model, it can incorporate personal information of a particular user via model adaptation techniques to make custom predictions. In light of the cognitive load for annotating music emotion, we cannot assume that a user is willing to provide a sufficient amount of personal annotations at once to make the system reach an acceptable performance level. On the contrary, a user may provide annotations sporadically in different listening sessions. An online learning strategy is therefore desirable. When the annotations of a target user are scarce, a good online learning method needs to prevent over-fitting to the personal data in order to keep certain model generalizability.

Motivated by the GMM-universal background model developed for speaker verification, we can first treat the affective GMM learned from broad subjects as a *background model*, and then employ, for example, maximum likelihood linear regression (MLLR) [1] or maximum a posteriori (MAP) methods [7] to update the parameters of the background model in an online fashion using the personal annotations. The resulting *personalized model* should find a good trade-off between the target user's annotations and the background model.



(a) Point-based (b) Trajectory-based

Fig. 2. The stress-sensitive user interface for emotion-based music retrieval. Users can (a) specify a point or (b) draw a trajectory, while specifying the variance with different levels of duration.

Emotion-based Music Retrieval

The VA space creates a straightforward visualization means for music collection browsing [9]. Moreover, it offers a ready canvas for music retrieval through the specification of a point in the emotion space. For example, users can retrieve music clips of certain emotions without specifying the song titles or artist names, or draw a trajectory to indicate the desired emotion changes across a list of songs (e.g. from tenderness to anger) [11]. An interesting retrieval approach enabled by AEG is to use a Gaussian-based query by specifying the desired variances (or the confidence level at the center point) of emotion by pressing a point in the VA space with different levels of duration or strength. As Figure 2 illustrates, the variance of the Gaussian gets smaller as one increases the duration or strength of pressing. Larger variances indicate less specific emotion around the center point. After specifying the size of a circular variance shape, one can even pinch fingers to adjust the variance shape. For a trajectory-based query input, similarly, the corresponding variances are determined according to the dynamic speed when drawing the trajectory. Fast speed corresponds to a less specific query and the system will return pieces whose variances of emotion are larger. If songs with more specific emotions are desirable, one can slow down the speed when drawing the trajectory. Such queries can be handled by AEG by various ways, as shown in [5].

Emotion-based Music Video Generation

Nowadays, everyone can easily create a video sequence by a consumer camcorder and broadcast it over the Internet through video sharing websites such as YouTube. To enhance the entertaining and aesthetic qualities of the video sequences, it is useful to accompany a video sequence with a piece of music that goes well together. For example, people like to accompany sports video by exciting music, and cheerful music for home video. We can formulate machine-based automatic generation of music videos as a cross-modality retrieval problem and tackle it based on an extension of AEG. Specifically, this can be done by jointly learning the tripartite relationship among music, video, and emotion from an emotion-annotated corpus of music videos. Figure 3 is an illustration of this Acoustic-visual Emotion Gaussians (AVEG) model. For a music clip (or a video sequence), the AVEG model is applied to predict its emotion distribution in the VA space from the corresponding

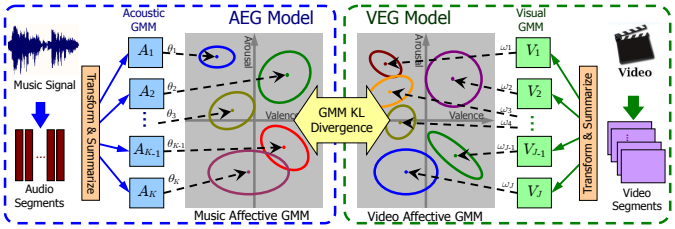


Fig. 3. System diagram of the Acousticvisual Emotion Gaussians (AVEG) model for emotion-based music video generation.

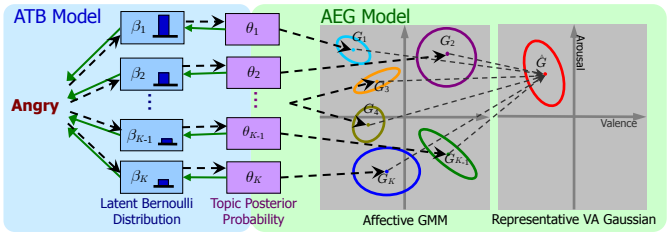


Fig. 4. Illustration of the generation flow between tag-based and VA-based emotion semantics of music. Two component models, namely Acoustic Tag Bernoullis (ATB) and AEG, are shown in the left and right panels, respectively. The representative VA Gaussian of a tag can be generated by following the black dashed arrows.

low-level acoustic (resp. visual) features. Finally, music and video are matched by measuring the distance between the two corresponding emotion distributions, based on a distance measure such as the KL divergence. This model won the first prize in the 2012 ACM SIGMM Multimedia Grand Challenge. Please refer to [4] for more details.

Connecting Emotion Dimensions and Categories

In addition to describing emotions by dimensions, emotions can also be described in terms of discrete labels, or tags. While the dimensional approach offers a simple means for constructing a 2-D user interface, the categorical approach offers an atomic description of music that is easy to be incorporated into conventional text-based retrieval systems. Being two extreme scenarios (discrete/continuous), the two approaches actually share a unified goal of understanding the emotion semantics of music. As the two approaches are functionally complementary, it is therefore interesting to investigate the relationship between them and combine their advantages to enhance the performance of emotion-based music retrieval systems. For example, as a novice user may be unfamiliar with the essence of the valence and activation dimensions, it would be helpful to display emotion tags in the emotion space to give the user some cues. This can be achieved if we have the mapping between the emotion tag space and the VA space.

Based on AEG, we can unify the two semantic modalities under a unified probabilistic framework, as illustrated in Fig. 4. Specifically, we establish a probabilistic framework consisting of two component models, the *Acoustic Tag Bernoullis* (ATB) model and the AEG model, to computationally model the generative processes from acoustic features to the perceptions of an emotion tag and a pair of valence-activation values, respectively. The latent topics $\{z_k\}_{k=1}^{K}$ can act as a bridge between the two spaces, so that the ATB and AEG models can share and transit the semantic information to each other. The latent topics are learned directly from acoustic feature vectors, and thus the training datasets for learning the ATB and AEG models can be totally separate, relieving the requirement for a jointly-annotated dataset for the two emotion modalities. Interested readers are referred to [3] for details.

Conclusions

In this letter, we have presented the main ideas of a novel generative model called AEG for music emotion applications. We have also presented five applications of AEG, including MER, personalized MER, emotion-based music retrieval, emotion-based music video generation, and connecting emotion dimensions and categorical tags. As AEG is a generic framework, it can be easily extended to other multimedia data such as speech, image and video. It is hoped that the letter can bring more attention to affective-based applications for multimedia retrieval and recommendation.

References

[1] Y.-A. Chen, J.-C. Wang, Y.-H. Yang, and H.-H. Chen, "Linear regression-based adaptation of music emotion recognition models for personalization," in *Proc. IEEE ICASSP*, 2014, pp. 2149–2153.

[2] J. A. Russell, "A circumplex model of affect," *J. Personality and Social Science*, vol. 39, no. 6, pp. 1161–1178, 1980.

[3] J.-C. Wang, Y.-H. Yang, K.-C. Chang, H.-M. Wang, and S.-K. Jeng, "Exploring the relationship between categorical and dimensional emotion semantics of music," in *Proc. ACM Int. Works. Music Information Retrieval with User-centered & Multimodal Strategies*, 2012, pp. 63–68.

[4] J.-C. Wang, Y.-H. Yang, I. Jhuo, Y.-Y. Lin, and H.-M. Wang, "The acousticvisual emotion Gaussians model for automatic generation of music video," in *Proc. ACM Multimedia*, 2012, pp. 1379–1380.

[5] J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Affective music information retrieval," in *Emotions and Personality in Personalized Services*, M. Tkalk̂iḱ̂ *et al.*, Eds. Springer, 2015.

[6] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "The acoustic emotion Gaussians model for emotion-based music annotation and retrieval," in *Proc. ACM Multimedia*, 2012, pp. 89–98.

[7] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Personalized music emotion recognition via model adaptation," in *Proc. APSIPA Annual Summit & Conference*, 2012.

[8] J.-C. Wang, Y.-H. Yang, H.-M. Wang, and S.-K. Jeng, "Modeling the affective content of music with a Gaussian mixture model," *IEEE Trans. Affective Computing*, 2015, in press.

[9] Y.-H. Yang and H. H. Chen, *Music Emotion Recognition*. CRC Press, 2011.

[10] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, 2012.

[11] Y.-H. Yang, Y.-C. Lin, H.-T. Cheng, and H. H. Chen, "Mr. Emo: Music retrieval in the emotion plane," in *Proc. ACM Multimedia*, 2008, pp. 1003–1004.

Post-Filter Using Modulation Spectrum as a Metric to Quantify Over-Smoothing Effects in Statistical Parametric Speech Synthesis

Shinnosuke Takamichi*†, Tomoki Toda*, Alan W Black† and Satoshi Nakamura*
* Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
E-mail: shinnosuke-t@is.naist.jp
† Language Technologies Institute, Carnegie Mellon University (CMU), U. S. A

***Abstract*—Recently, we have found a Modulation Spectrum (MS) is capable of sensitively quantifying over-smoothing effects, which is one of the main factors causing quality degradation in synthetic speech. This letter briefly describes our proposed approach based on a post-filter to modify the MS to alleviate the over-smoothing effects.**

I. Introduction

There is no doubt that speech-based systems have been changing our lifestyle. Speech synthesis techniques, such as Text-To-Speech (TTS) synthesis [1] or Voice Conversion (VC) [2], are one of the key technologies to develop these systems. In 1990s, the basic idea of the statistical parametric approaches to speech synthesis was originally proposed [3], [2], and recently many attempts, such as Hidden Markov Model (HMM) [4], Gaussian Mixture Model [5], kernel regression [6], and deep neural nets-based approaches [7], have been studied to deploy the speech-based systems.

One of the common issues of the statistical parametric speech synthesis is quality degradation of the synthetic speech [8], [9]. The synthetic speech parameters predicted from the input information are often overly smoothed, and this over-smoothing effect causes muffled sounds in synthetic speech. There have been various approaches to alleviate this over-smoothing effect, e.g., modification of the synthetic speech parameters to revise their features clearly different from those in natural speech parameters. However, this issue still remains to be addressed.

Recently, we have found a Modulation Spectrum (MS) of the synthetic speech parameters as a novel feature to sensitively quantify the over-smoothing effect and have proposed several post-filter based implementations to modify it to improve synthetic speech quality [10], [11], [12]. Our approach has the following merits.

Supported by previous work: The effectiveness of the MS have been reported on speech perception and speech recognition [13], [14]. Furthermore, the MS is also regarded as the extension of the Global Variance (GV) [5], [15] known as a conventional feature to quantify the over-smoothing effect.

Application to various speech components: Though some conventional post-filters, such as cepstral emphasis [16] or peak-to-valley emphasis [17], can be applied to only spectral features, our proposed post-filter is applicable to not only the spectral component but also excitation and duration components.

High portability: Because our proposed post-filter less depends on the speech synthesis procedure, it is available in various speech synthesizers. In **Section IV**, its effectiveness will be verified in HMM-based TTS [4] and GMM-based VC [15].

This paper briefly describes the concept of the MS-based post-filter. The post-filter is trained in advance using the natural and synthetic speech parameters, and then, it is applied to the synthetic speech parameters in a synthesis stage. Note that a part of our work has been implemented in HTS (HMM-based Speech Synthesis System) version 2.3 beta [18].

II. Statistical Parametric Speech Synthesis and Over-Smoothness Analysis

A. Speech Parameter Prediction from Input Information

In training, a specific statistical model, e.g., HMM or GMM, models a probability density function of the natural speech parameters corresponding to input information. The synthetic speech parameters are predicted based on the maximum likelihood criterion using the trained statistical models. The averaging process in these frameworks often excessively removes the detailed characteristics of the natural speech parameters, and causes the muffled effects in synthetic speech. Although the GV is well incorporated in the parameter generation stage [5], [15], its quality gains are limited.

B. Over-Smoothness Analysis Using Modulation Spectrum

Here, we analyze the natural and generated speech parameter sequences based on the MS. The MS is defined as the power spectrum of the parameter sequence; i.e., temporal fluctuation of the parameter sequence is decomposed into individual modulation frequency components and their power values are represented as the MS. Figure 1 illustrates the averaged MS of the natural ("natural") and generated ("HMM" and "HMM+GV") mel-cepstral coefficients in HMM-based TTS. We can observe that the generated MS is markedly degraded at the higher modulation frequency. This is because high temporal fluctuation observed in the natural speech parameter sequence is usually lost in the parameter generation

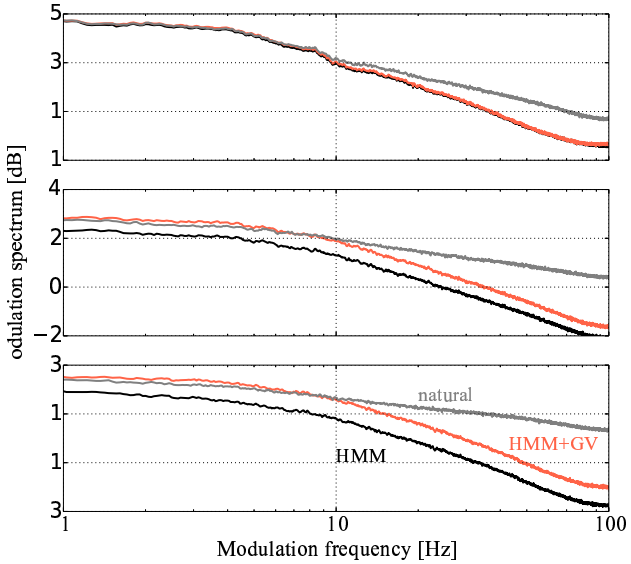


Fig. 1. Averaged modulation spectra of 1st, 9th, and 15th mel-cepstral coefficient sequences in HMM-based TTS.

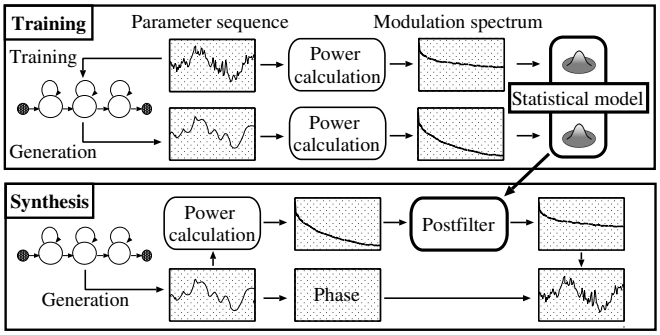


Fig. 2. A schematic diagram of the MS-based post-filter in the case of HMM-based TTS.

frameworks. We can also observe that the degradation tends to be larger in higher order of the mel-ceptral coefficients.

III. Modulation Spectrum-Based Post-Filter

Our proposed post-filter is designed to make the MS of the generated speech parameters close to that of the natural ones. Figure 2 shows the schematic diagram of the post-filter in the case of HMM-based TTS. The post-filter is automatically trained using natural and generated parameter sequences in the training data.

A. Basic Process

1) Training: The training stage estimates the statistics of the natural and generated MS. Let $s_d(f)$ be the f-th MS of the d-th dimension of the parameter sequence. The mean $\mu_{d,f}^{(\mathrm{N})}$ and variance $\left(\sigma_{d,f}^{(\mathrm{N})}\right)^2$ of $s_d(f)$ are estimated using the natural parameter sequences. The mean $\mu_{d,f}^{(\mathrm{G})}$ and variance $\left(\sigma_{d,f}^{(\mathrm{G})}\right)^2$ are estimated in the same manner using the parameter sequences generated by a statistical parametric speech synthesis framework.

2) Conversion: The following filter is applied to the generated speech parameter sequence:

$$s'_d(f) = (1-k)s_d(f) + k\left[\frac{\sigma_{d,f}^{(\mathrm{N})}}{\sigma_{d,f}^{(\mathrm{G})}}\left(s_d(f)-\mu_{d,f}^{(\mathrm{G})}\right)+\mu_{d,f}^{(\mathrm{N})}\right], \quad (1)$$

where k is a post-filter emphasis coefficient valued between 0 and 1. The finally filtered parameter trajectory is calculated from the filtered MS and phase components extracted from the parameter sequence before filtering.

B. Various Implementations

1) Application to F_0 and Duration: The post-filtering process is available in various speech features. In the case of the F_0 contour [10], the post-filter cannot be directly applied because the observed F_0 contour is not a continuous sequence. To address this issue, continuous F_0 contour calculated from the observed F_0 contour is firstly filtered, and then, the U/V region is restored. For duration [12] in HMM-based TTS, the post-filter is effectively applied to not HMM-state but phoneme duration. The modified HMM-state is calculated by maximizing the duration likelihood given the filtered phoneme duration.

2) Segment-Level Filtering: Because the original post-filter [10] performed the utterance-level conversion, the MS calculation is not accurately calculated when the length of an utterance to be synthesized is longer than the previously determined DFT length. To address this problem, we proposed another post-filter implementation that modifies the MS of the segmented parameter sequence [12]. The filtered parameter sequence is generated by overlapping and adding the segmented sequence.

IV. Experimental Evaluation

We conducted several subjective evaluations in order to confirm the quality gain by the post-filter. The experimental condition is described in [10], [11], [12]. Figures. 3-6 illustrate the results of the preference AB test on speech quality. The error bar is 95% confidence interval. “HMM” or “GMM” performed the ML-based parameter generation [19], [15], “*+GV” considered the GV in the parameter generation [5], [15], “*+MSPF” applied the post-filter. The post-filter emphasis coefficients were determined by objective evaluations [10], [11], [12]. In these figures, the post-filter have achieved better-quality than the conventional approaches, and the quality gain is the largest in the spectral features. Even when we consider the GV in parameter generation, the additional quality gain is observed by the post-filter. This is because the MS compensation by the GV-based approaches is insufficient as shown in Figure 1.

Figure 7 illustrates an example of the mel-ceptral coefficient sequences. We can see that the post-filter makes the sequence

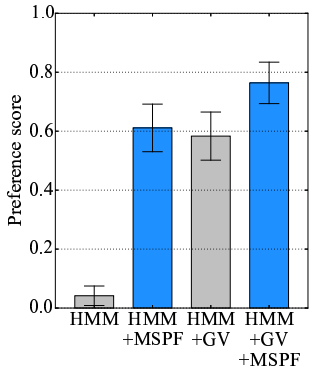


Fig. 3. Preference scores (mel-cepstrum in HMM-based TTS)

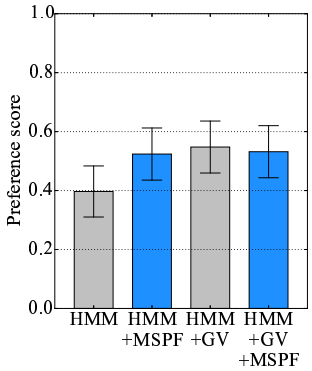


Fig. 4. Preference scores (F_0 contour in HMM-based TTS)

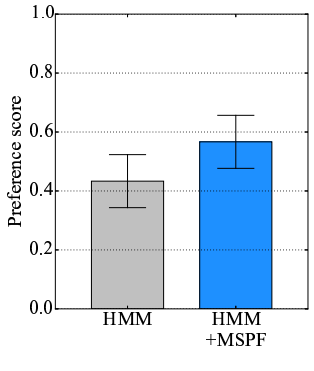


Fig. 5. Preference scores (duration in HMM-based TTS)

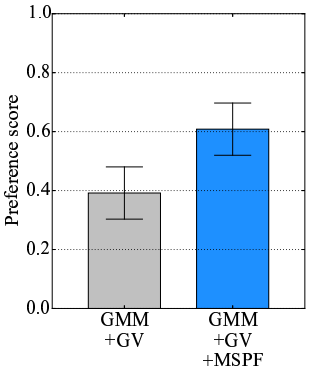


Fig. 6. Preference scores (mel-cepstrum in GMM-based VC)

fluctuated, and and such a fluctuation tends to be larger in higher order mel-cepstral coefficient.

V. Conclusion

This letter presented our approach using Modulation Spectrum (MS) for high-quality statistical parametric speech synthesis. The MS represents the fluctuation of the speech parameter sequence, and the MS-based post-filter improves the synthetic speech quality by recovering the MS of the generated parameter sequence.

Acknowledgements: Part of this work was supported by JSPS KAKENHI Grant Number 26280060 and Grant-in-Aid for JSPS Fellows Grant Number $26 \cdot 10354$, and part of this work was executed under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation."

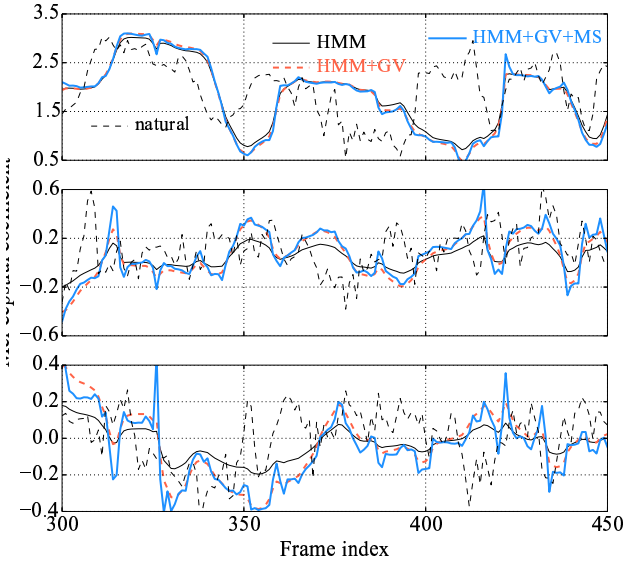


Fig. 7. Examples of 1st, 9th, and 15th mel-cepstral coefficient sequences in HMM-based TTS.

References

[1] Y. Sagisaka. Speech synthesis by rule using an optimal selection of non-uniform synthesis units. In *Proc. ICASSP*, pp. 679–682, New York, U.S.A, Apr. 1988.

[2] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. Speech and Audio Processing*, Vol. 6, No. 2, pp. 131–142, Mar. 1988.

[3] K. Tokuda, T. Kobayashi, and S. Imai. Speech parameter generation from HMM using dynamic features. In *Proc. ICASSP*, pp. 660–663, Detroit, U.S.A, May 1995.

[4] K. Tokuda, Y Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura. Speech synthesis based on hidden markov models. *Proceedings of the IEEE*, Vol. 101, No. 5, pp. 1234–1252, 2013.

[5] T. Toda and K. Tokuda. A speech parameter generation algorithm considering global variance for HMM-based speech synthesis. *IEICE Trans.*, Vol. E90-D, No. 5, pp. 816–824, 2007.

[6] E. Helander, T. Virtanen H. Silen, and M. Gabbouj. Voice conversion using dynamic kernel partial least squares regression. *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 3, pp. 806–817, Mar. 2012.

[7] H. Zen and A. Senior. Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *Proc. ICASSP*, pp. 3872–3876, Florence, Italy, May 2014.

[8] S. King and V. Karaiskos. The blizzard challenge 2011. In *Proc. Blizzard Challenge workshop*, Turin, Italy, Sept. 2011.

[9] Y. Stylianou. Voice transformation: A survey. In *Proc. ICASSP*, pp. 3585–3588, Taipei, Taiwan, Apr. 2009.

[10] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. A post-filter to modify modulation spectrum in HMM-based speech synthesis. In *Proc. ICASSP*, pp. 290–294, Florence, Italy, May 2014.

[11] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modulation spectrum-based post-filter for gmm-based voice conversion. In *Proc. APSIPA ASC*, Siem Reap, Cambodia, Dec. 2014.

[12] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura. Modified modulation spectrum-based post-filter for hmm-based speech synthesis. In *Proc. GlobalSIP*, pp. 710–714, Atlanta, United States, Decc. 2014.

[13] R. Drullman, J .M. Festen, and R. Plomp. Effect of reducing slow temporal modulations on speech reception. *J. Acoust. Soc. of America*, Vol. 95, pp. 2670–2680, 1994.

[14] S. Thomas, S. Ganapathy, and H. Hermansky. Phoneme recgnition usng spectral envelop and modulation frequency features. In *Proc. ICASSP*, pp. 4453–4456, Taipei, Taiwan, April 2009.

[15] T. Toda, A. W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Incorporation of mixed excitation model and postfilter into HMM-based text-to-speech synthesis. *IEICE Trans. Inf. Syst.*, Vol. J87-D-II, No. 8, pp. 1563–1571, 2004.

[17] F. Eyben and Y. Agiomyrgiannakis. A frequency-weighted post-filtering transform for compensation of the over-smoothing effect in HMM-based speech synthesis. In *Proc. ICASSP*, pp. 275–279, Florence, Italy, May 2014.

[18] HMM-based speech synthesis system (HTS) http://hts.sp.nitech.ac.jp/.

[19] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

Interference Suppression Schemes for Radio over Fiber Simultaneously Transmitted with 10 Gbps On-Off Keying

Yuya Kaneko, Takeshi Higashino, and Minoru Okada
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0192 JAPAN
Email: kaneko.yuya.kr2@is.naist.jp

I. Introduction

The radio over fiber (RoF) offers small base stations, centralized operation for heterogeneous wireless service, cooperative distributed antenna system, and large transmission capacity. The typical RoF link is configured by the intensity modulation/direct detection (IM/DD) link. Radio frequency (RF) signal directly modulates the laser diode (LD) or externally modulates the light source. The RF signal is detected by a photodetector (PD) after the transmission over optical fiber channel. Meanwhile, optical On-Off Keying (OOK) modulation for baseband (BB) transmission is employed in 10 Gbps Ethernet PHY layer. The 10 Gbps Ethernet infrastructure is widely used in in-building local area network (LAN). The optical fiber communication technology is employed as various broadband transmission such as Ethernet, mobile backhaul, and mobile fronthaul. However, these provide single communication service. The realization for providing multiservice motivates the simultaneous transmission for baseband and RF signals[1]-[4]. It enables independent different types of networks to share infrastructure and realizes quick and cost effective implementation.

In previous papers[1]-[3], the frequency band of RF signal is higher than the main lobe of the OOK spectrum. Since the frequency channels of major cellular system and Wireless LAN (WLAN) such as the Long-Term Evolution (LTE) and the WiFi are assigned at lower than 10 GHz, investigation for coexistence of the radio systems with frequency lower than 10 GHz and 10 Gbps optical BB signal is required. Chen et al.[4] investigated simultaneous transmission of 10 Gbps OOK signal and RF signal at 2 GHz. The 8B10B channel coding is employed in their proposal in order to make spectrum notch with specific interval. The RF signal can be allocated at the notch frequency band due to suppressed interference. This scheme cannot be applicable to standard ethernet because 64B66B channel coding is employed, there is no spectrum notch over the OOK spectrum. Previous study[5] proposed another configuration of RoF link with optical OOK signal. The proposed system is called as radio on optical OOK (RoOOOK). The RoOOOK is easy to construct over the existing fiber link, however, the experiment reveals that simultaneously transmitted OOK signal interferes with RF signal[5].

In this paper, interference suppression schemes in the RoOOOK are proposed. At first, a theoretical analysis of its power spectrum is presented and it is compared with the experiment. In the analysis, the stochastic process of optical OOK stream is re-modulated by the RF signal. Signal to noise power ratio (SNR) and error vector magnitude (EVM) of the RF signal are derived. An experimental evaluation using 1.9 GHz RF signal is conducted to investigate coexistence with 10 Gbps OOK signal. Experimental results agree with the theory. The experiment shows that the simultaneous transmission of the RF and BB signals without serious performance degradation is possible without any interference suppression. However, interference is still a problem, therefore we propose interference suppression schemes to improve EVM and widen dynamic range of the system. The improvement of SNR and dynamic range is discussed from theoretical analysis and computer simulation.

II. System description

Fig. 1 shows the configuration of RoOOOK using 10 Gbps Ethernet. A pair of 10 Gbps Ethernet switches is connected by optical fibers. The transmitting optical stream is re-used as a carrier for RF modulation. The OOK is employed as a modulation format of 10 Gbps Ethernet. The amplitude of RF signal is mapped onto the intencity of the optical pulse using an external optical modulator (EOM). At the receiving side, optical signal is divided into two branches using optical coupler (OC). One branch is led to the OOK demodulator in the Ethernet receiver. The other signal is fed into the optic-to-electric (O/E) converter. Since the obtained photo current is still waveform of pulsed RF signal, the electric band-pass filter (BPF) is used to regenerate the original RF signal.

In the viewpoint of RF signal transmission, it can be considered as the band-pass sampling and regeneration, because the RF signal is sampled by the random pulse sequence. Therefore, the spectrum analysis with stochastic process can be applicable.

III. Power spectrum analysis

Fig. 2 shows the external re-modulation of the OOK signal. Consider $p(t)$ as the rectangular pulse waveform taking 1 or 0 with its duration of $1/f_p$, $s_r(t)$ as the RF signal, B as the amplitude of pulses, and $0 \leq m \leq 1$ as the modulation index. The re-modulated OOK signal is expressed as,

$$v(t) = \{1 + ms_r(t)\}Bp(t). \quad (1)$$

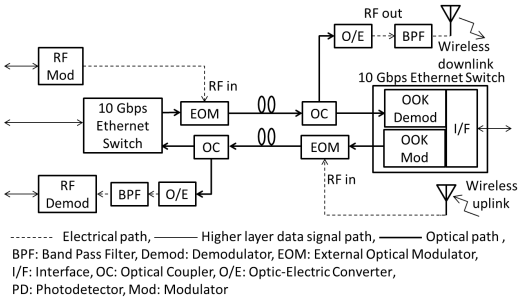


Fig. 1: Radio on optical OOK (RoOOOK) using 10 Gbps Ethernet.

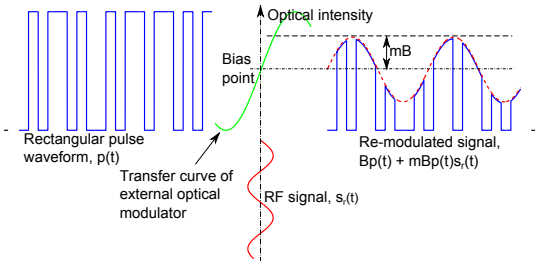


Fig. 2: OOK signal externally re-modulated by RF signal.

The autocorrelation function of $v(t)$, $R_v(\tau)$, is expressed as,

$$R_v(\tau) = B^2 R_p(\tau) + m^2 B^2 R_p(\tau) R_s(\tau), \quad (2)$$

where τ is the time lag, $R_p(\tau)$ is the autocorrelation function of $p(t)$, and $R_s(\tau)$ is the autocorrelation function of $s_r(t)$. Assumiing that the OOK signal is non return-to-zero OOK bit sequence, any bits are statistically independent, and the probability of occurrence for mark ($p(t) = 1$) is ρ, then On the basis of Wiener-Khinchin theorem, the Fourier transform of (2), $G_v(f) = \mathcal{F}[R_v(\tau)]$, is the power spectral density (PSD) of the transferred signal.

$$G_v(f) = B^2\rho^2\delta(f) + \frac{\rho(1-\rho)}{f_p} B^2 \mathrm{sinc}^2\left(\frac{f}{f_p}\right) + m^2 B^2 \Big\{ \rho^2 G_s(f) + \frac{\rho(1-\rho)}{f_p} \mathrm{sinc}^2\left(\frac{f}{f_p}\right) * G_s(f) \Big\}, \quad (3)$$

where $G_s(f)$ is the Fourier transform of $R_s(\tau)$ and the operator $*$ means convolution. The first term on the right-hand side of (3) is the DC component. The second term is the frequency component of the OOK signal. The third term is the RF signal. The fourth term is the noise caused by the sampling of the RF signal by the OOK signal thus we call this term alias. Fig. 3 shows an example of PSD.

The peak amplitude of $s_r(t)$ is normalized by 1 and the power of RF signal is considered as a combination of m, B, and S_s which is the power of $G_s(f)$. S_s equals $1/2$ when $s_r(t)$ is a sinusoidal wave. Assuming that the occupied bandwidth of $G_s(f)$ is the center frequency f_r and the bandwidth W, S which is the signal power after through BPF is,

$$S = \left(\int_{-f_r - W/2}^{-f_r + W/2} + \int_{f_r - W/2}^{f_r + W/2} \right) G_v(f) \mathrm{d}f. \quad (4)$$

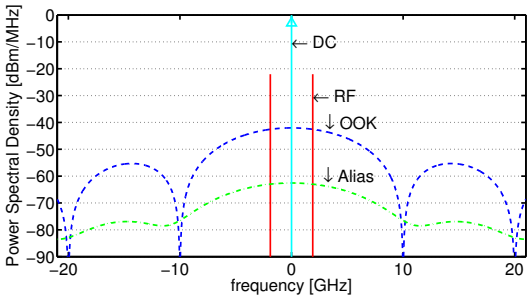


Fig. 3: Power spectral density of RoOOOK ($f_p = 10$ Gbps, $f_r = 1.9$ GHz, $W = 384$ [kHz], $B = 0.05$ volt, and $m = 0.1$).

TABLE I: Specification of the equipments.

XFP specification	
PHY Std.	10GBASE-ER
Tx. power	0 dBm
Rx. sensitivity	-16 dBm
RF specification	
Carrier freq.	1.9 GHz
Bit rate	384 [kbps]
Modulation	$\pi/4$ QPSK
LPF	RNYQ, α = 0.5
Optical system specification	
λ	1550 nm
fiber	SMF
V_π	5 volt

The SNR of the RF signal, γ, is expressed as a function of m and ρ,

$$\gamma(m,\rho) = \frac{m^2 B^2 \rho^2 S_s}{S - m^2 B^2 \rho^2 S_s + N_0 W} \frac{W}{f_s}, \quad (5)$$

where f_s is the symbol rate of the RF signal, and N_0 is the one-sided noise power spectral density.

IV. Experiment

Tab. I shows the specification of the equipments. Fig. 4 shows the experimental setup. IM/DD RoF with external modulation (EM) using continuous wave light source is compared with the proposal. The average optical power of the LD is 3 dB higher than the output of the optical Ethernet switch. This setting equalizes the instantaneous maximum incident optical power to the Mach-Zehnder modulator which is used as an EOM. The measured total optical power loss between Ethernet switches including connector loss, fiber loss, insertion loss, biasing of modulator, and splitting loss is 11 dB.

Fig. 5 shows the PSDs of both RoF with EM and RoOOOK. The RF signal power in RoOOOK is less than RoF with EM by 6 dB. It is caused by the 3 dB difference of the incident optical power to EOM. The spectral components of OOK interfere with the RF signal. It corresponds to the result of the theoretical analysis shown in (3) and Fig. 3.

Theoretical EVM is obtained from the relation of average EVM and SNR[6], $EVM \simeq 1/\sqrt{SNR} = 1/\sqrt{\gamma(m,\rho)}$. A typical value of ρ is considered as $1/2$ because the occurrence for mark ($p(t) = 1$) and space ($p(t) = 0$) are equiprobable. The incident RF power to EOM is related to m by depending

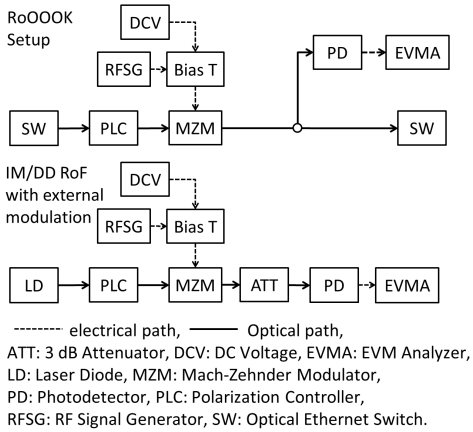


Fig. 4: Experimental setup.

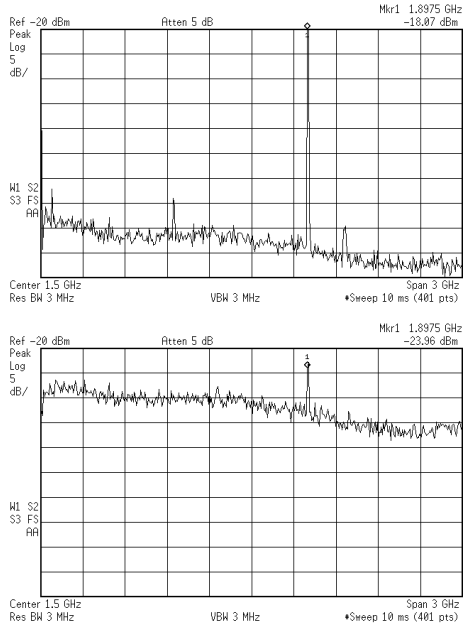


Fig. 5: Power spectral density of IM-DD RoF with external modulation (top) and RoOOOK (bottom), RBW $= 3$ MHz, RLV $= -20$ dBm, 5 dB/div, RF in $= 0$ dBm.

on $m = \sqrt{2}V_{rms}/V_\pi/2$, where V_{rms} is the root mean square voltage of RF signal and V_π is the half-wave voltage of the EOM. Fig. 6 shows a comparison between the theory and the experiment in EVM. The theoretical curve corresponds to the experimental result. Although deep superimposing can improve EVM, it yields service outage in Ethernet link. When incident RF power exceeds over 6 dBm, the throughput of Ethernet link significantly decreases and the link is in a service outage. The required EVM for RF signal and the service outage of 10 Gbps Ethernet link can define the dynamic range of the system. In this case, required EVM is 12.5 %, then the minimal required incident RF power is found to be -6 dBm. Therefore, the dynamic range is 12 dB between -6 dBm and 6 dBm of incident RF power.

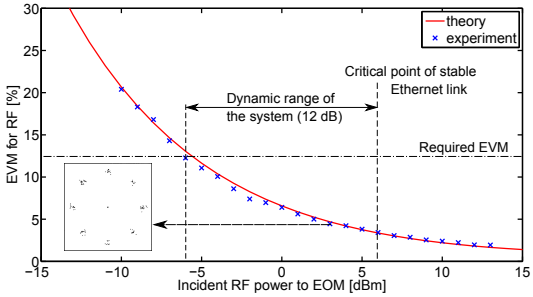


Fig. 6: Relationship between EVM and incident RF power to EOM.

V. Interference Suppression Using Rectification

In this section, we propose interference suppression schemes using rectification. The proposal outperforms the conventional in terms of EVM.

A. Description of proposed scheme

As shown in Figs. 3 and 5, a spectral component of OOK interferes with an RF signal and deteriorates SNR. Half-wave rectification can reduce the spectral component of OOK. Fig. 7 shows a proposed interference suppression scheme using half-wave rectification. The DC component is blocked, then the obtained signal is rectified. The half-wave rectified signal is expressed as $m_h = 2m$ and $B_h = B/2$ for m and B in (1), respectively. Therefore, the half-wave rectification is equivalent to doubling the modulation index.

As the development of this scheme, biased half-wave rectification is more effective when the modulation index is knowable at the receiver side. Fig. 8 shows a proposed interference suppression scheme using biased half-wave rectification. The received signal is biased by $(1-m)B$ and then it is half-wave rectified. The half-wave rectified signal is expressed as $m_h = 2m$ and $B_h = B/2$ for m and B in (1), respectively. The biased half-wave rectification is equivalent to making the modulation index 1.

Undesirable physical properties of rectifier (e.g., reverse recovery time and nonlinearity of diode) are ignored in this paper.

B. Simulation result

A theoretical curve of EVM versus incident RF power to EOM is drawn in the same way as section IV. Figs. 9 and 10 show theoretical EVM curves and simulation results of conventional RoOOOK and proposal with noise power densities $N_0 = -60$ and -70 dBm/MHz, respectively. Typically, N_0 is determined by the thermal noise of a receiver. These figures show improvement of EVM. Fig. 9 shows that the half-wave rectification improves EVM by about 6 dB and the biased half-wave rectification improves EVM by about 20 dB. Therefore, the required minimal incident RF power can be reduced by 6 dB or 20 dB compared with the conventional. This power reduction contributes improvement for the system dynamic range.

The improvement effect due to use of the half-wave rectification in Fig. 10 is not different from the result of Fig. 9.

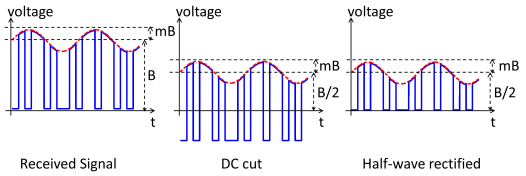


Fig. 7: The waveform of half-wave rectified signal.

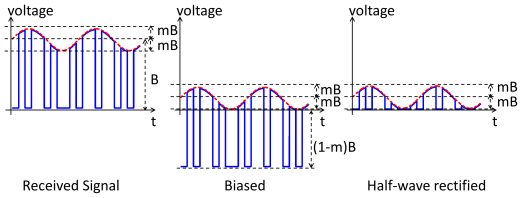


Fig. 8: The waveform of biased half-wave rectified signal.

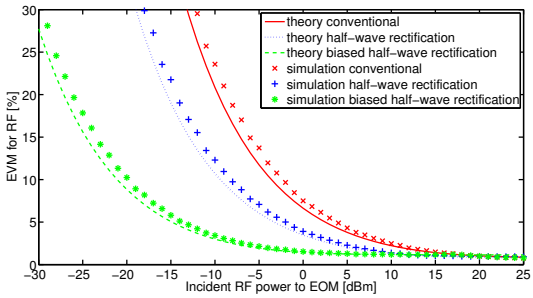


Fig. 9: The simulation results of relationship between EVM and incident RF power to EOM at $N_0 = -60$ dBm/MHz.

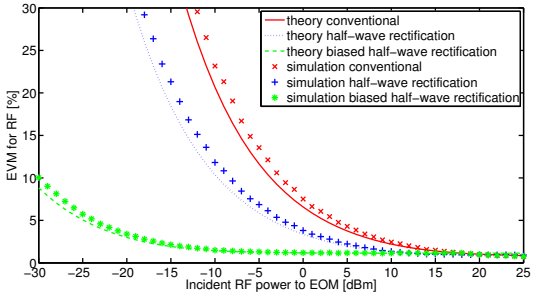


Fig. 10: The simulation results of relationship between EVM and incident RF power to EOM at $N_0 = -70$ dBm/MHz.

In contrast, The improvement effect due to use of the biased half-wave rectification in Fig. 10 is better than the result of Fig. 9. In the conventional RoOOOK and the half-wave rectified RoOOOK, the dominant noise source is the spectral component of OOK signal, because it is generally higher than the thermal noise. Therefore, the difference of N_0 makes no change of EVM curve. In the biased half-wave rectified RoOOOK, the dominant noise source changes depending on the modulation index m which relates to incident RF power. In a range of low incident RF power which means low m, the dominant noise is the thermal noise. By the use of the biased half-wave rectification, the required minimal incident RF power is determined by the thermal noise of the receiver. In a range of high incident RF power which means high m, the dominant noise is the spectrum of OOK component in each case. Therefore, the EVM curve has the lower limit. However, the lower limit presented in Fig. 9 is sufficiently low to satisfy the requirement. The proposed schemes improve EVM and widen the dynamic range.

VI. Conclusion

This paper proposed interference suppression schemes for RoF simultaneously transmitted with 10 Gbps optical OOK. After theoretical power spectrum analysis, theoretical EVM performance is compared with the experiment on radio signal transmission over 10 Gbps optical Ethernet link. The theoretical EVM corresponds to the experimental results. The external re-modulation of the optical OOK signal yields impact on BB and RF resulting in an service outage for 10 Gbps Ethernet link, and less dynamic range compared with an independent RoF link. Experimental results show the both two signals can coexist and share single fiber channel with the dynamic range of 12 dB. To improve EVM and system dynamic range, interference suppression schemes are proposed. The improvement is discussed from the theoretical analysis and simulation results. They imply that the proposed scheme using biased half-wave rectification improves the dynamic range by about 20 dB. The proposed schemes are expected to be implemented by analog electrical circuits. An empirical proof of the proposed scheme is a future work.

References

[1] C.T. Lin, et al., "Hybrid Optical Access Network Integrating Fiber-to-the-Home and Radio-Over- Fiber Systems," IEEE Photonics Technology Letters, Vol.**19**, no.8, Apr.2007.

[2] C.W. Chow, C.H. Yeh, C.H. Wang, F.Y. Shih and S. Chi, "Signal-Remodulated Wired/Wireless Access Using Reflective Semiconductor Optical Amplifier With Wireless Signal Broadcast," Photonics Technology Letters, IEEE , Vol.**21**, no.19, pp.1459,1461, Oct. 2009.

[3] S. Yaakob, et al., "Minimal optimization technique for radio over fiber WLAN transmission in IM-DD optical link," Microwave and Optical Technology Letters, Vol.**52**, issue 4, pp.812-815, 2010.

[4] C. Chen, R. Penty, I. White, and M. Crisp, "Transmission of Simultaneous 10Gb/s Ethernet and Radio-over-Fiber Transmission using In-band Coding," OFC/NFOEC 2013, OSA Technical Digest (online) (Optical Society of America, 2013), paper OM3D.1, 2013.

[5] Y. Kaneko, T. Higashino, and M. Okada, "An Empirical Performance Evaluation of Radio on Optical OOK System," IEICE Tech. Rep., vol. 113, no. 397, MWP2013-63, pp. 51-56, Jan. 2014.

[6] R. A. Shafik, S. Rahman, R. Islam, and N. S. Ashraf, "On the Error Vector Magnitude as a Performance Metric and Comparative Analysis," International Conference on Emerging Technologies (ICET), pp.27-31, 2006.

SEAME: South-East Asia Mandarin-English Speech Database

Ho Thi Nga, Chng Eng Siong, Haizhou Li

In this newsletter, we introduce the SEAME (South-East Asia Mandarin-English) database, the first publicly released Mandarin-English code-switching speech database for automatic speech recognition (ASR) research. The database was jointly developed by the Speech Groups in Temasek Laboratories@Nanyang Technological University, Singapore and in Universiti Sains Malaysia, Malaysia, and will be released by LDC (LDC2015S04) in Q2, 2015 [5]. SEAME includes 192 hours of Mandarin-English code-switching and mono-lingual utterances in conversation and interview settings.

What is code-switching speech?

Code-switching occurs when a speaker alternates between two or more languages in a conversation. It has become very common in bilingual societies, people code-switch to convey a thought, and to fit in a social context among many other reasons [1]. In Singapore and Malaysia, the population consists of multiple races who can speak their respective native languages well along with the common language, English. Hence, it is quite common to have speakers embedding English words or phrases when they are speaking their own native languages and vice versa.

Developing a code-switching ASR system is considerably more difficult than a mono-lingual one. This is because the code-switching system will need to accommodate within one sentence the various languages pronunciation and language model simultaneously. Furthermore, there are only very few code-switching corpora available publicly, and are small in scale [1].

Why SEAME corpus?

The SEAME corpus is designed to support many aspects of code-switching studies related to ASR as follows, among others.

- Pronunciation modelling
- Language modelling
- Acoustic modelling
- Automatic speech recognition
- Language identification [2]
- Language turn detection [3][4]

Database description

SEAME (LDC2015S04) consists of 192 hours of speech, of which 63 hours are manually labelled with word level transcription. Most of the 63 hours of speech are code-switching utterances. The un-transcribed speech data, mostly mono-lingual segments, will be released in the future.

The voices are recorded in two speaking styles: conversational (about 92 hours) and interview (about 100 hours). In conversational setting, the voices from both interlocutors were recorded, while in interview setting, the interviewee's voice was recorded. The corpus

was recorded in two sites (Singapore and Malaysia) with the same recording setup during 2009 to 2010, with a close-talk microphone in a quiet environment. Each recording session is about one hour long for either interview or conversation. There are 156 distinct speakers. They are balanced in gender with age between 19 and 33.

The transcriptions are given with sentence boundary. In addition to lexical words, the discourse particles are also provided in either Mandarin or English, for example [啊], [喔] in Mandarin and [ah], [oh] in English. The other meta-information when they occur, such as the appearance of other languages or non-speech events, is also labelled.

Where to get the database?

Visit LDC website: https://catalog.ldc.upenn.edu/LDC2015S04. The database will be available on the website from 15 April 2015 onward.

References

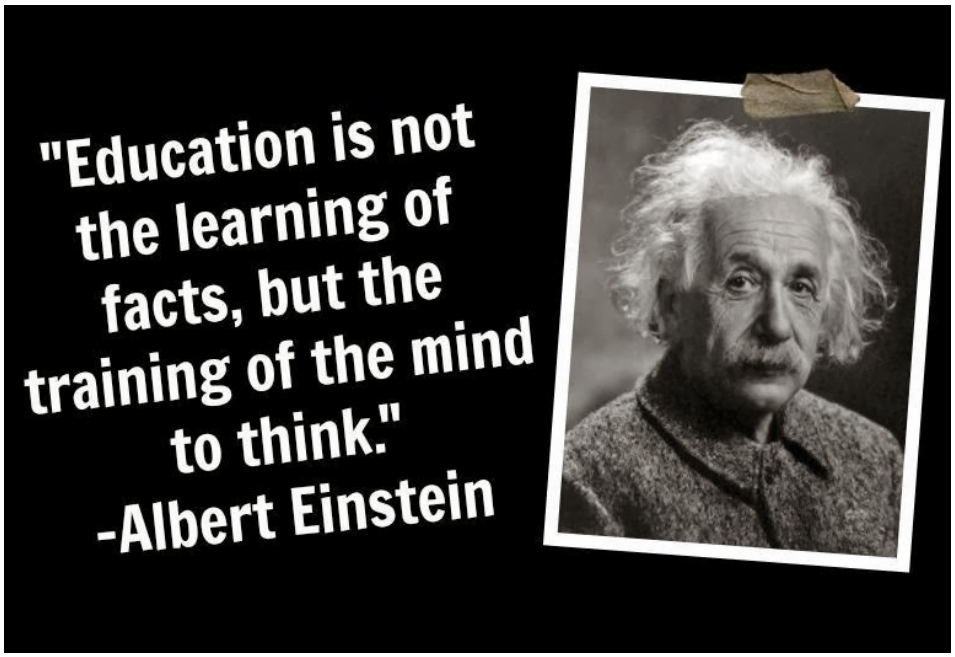
[1] Dau-Cheng Lyu, Tien Ping Tan, Engsiong Chng, Haizhou Li: SEAME: a Mandarin-English code-switching speech corpus in south-east asia. INTERSPEECH 2010: 1986-1989

[2] Ngoc Thang Vu, Dau-Cheng Lyu, Jochen Weiner, Dominic Telaar, Tim Schlippe, Fabian Blaicher, Engsiong Chng, Tanja Schultz, Haizhou Li: A first speech recognition system for Mandarin-English code-switch conversational speech. ICASSP 2012: 4889-4892

[3] Dau-Cheng Lyu, Eng-siong Chng, Haizhou Li: Language diarization for code-switch conversational speech with pronunciation dictionary adaptation. ChinaSIP 2013

[4] Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li: An Analysis of a Mandarin-English Code-switching Speech Corpus: SEAME. OCOCOSDA 2010.

[5] https://catalog.ldc.upenn.edu/LDC2015S04



CAMBRIDGE JOURNALS

APSIPA
TRANSACTIONS ON
SIGNAL AND INFORMATION
PROCESSING

Recent Articles:

- Lossless contour coding using elastic curves in multiview video plus depth; *Marco Calemme, Marco Cagnazzo and Beatrice Pesquet-Popescu*
- An overview of directivity control methods of the parametric array loudspeaker; *Chuang Shi, Yoshinobu Kajikawa and Woon-Seng Gan*
- Discriminating multiple JPEG compressions using first digit features; *Simone Milani, Marco Tagliasacchi and Stefano Tubaro*
- Digital acoustics: processing wave fields in space and time using DSP tools; *Francisco Pinto, Mihailo Kolundžija and Martin Vetterli*

Read all these articles at: journals.cambridge.org/sip/apsipa

Benefits of publishing

- Free color, no overlength page charges (flat $600 processing fee on accepted papers prior to publication)
- Broad range of topics covering both traditional signal processing applications and methods, as well as emerging areas and concepts
- Open access: free, permanent, worldwide access to your article
- Rapid publication: continuous publication of articles as soon as they are accepted
- Peer reviewed by international experts, first review response available in less than 2 months for full length papers (faster turn-around time for EWPs)
- State-of-the-art online hosting
- Papers indexed by Scopus: Indexing approved in May 2014
- Papers linked on Google Scholar; http://goo.gl/xWHRsf

Submit your paper online at: mc.manuscriptcentral.com/apsipa

CAMBRIDGE UNIVERSITY PRESS

Call for Proposals for APSIPA 2015 Tutorial Sessions

You are cordially invited to submit your proposals for 2015 APSIPA ASC tutorial sessions. We would like to enrich APSIPA conference with diverse knowledge through a set of tutorial sessions of 3 hours each. The tutorials are looked at as stepping stones to new researchers to undertake research in a certain field, boost the knowledge of experienced researchers, or direct the attention of the research community to an important field. The topics could be in any field within, but not limited to, the APSIPA themes:

- **Biomedical/Biological Signals and Systems**
- **Circuits, Design and Implementation**
- **Information Processing Theory and Applications**
- **Image, Video, and Multimedia**
- **Speech, Language, and Audio**
- **Wireless Communications and Networking**

As a reward APSIPA conference organisers dedicated an honorarium of $US1000 paid as a lump sum to the presenters of each tutorial session.

Please email your interest with the following information to the APSIPA Tutorial Sessions Co-Chair Waleed H. Abdulla w.abdulla@auckland.ac.nz

1. Full name(s) of the tutorial presenter (s)
2. Affiliation(s) (University, Organisation, Institutes)
3. Contact details (Postal Address, Phone, Fax, E-mail)
4. Title and abstract of the tutorial
5. Resume

IMPORTANT DATES

Submission of Proposals for Tutorial Sessions	May 8, 2015
Notification of Proposals Acceptance:	June 8, 2015
Tutorial Session Date	December 16, 2015

Looking forward to receiving your valuable proposals

Tutorial Co-Chairs: Waleed Abdulla, Woon-Seng Gan, Wing-Kuen Ling, Lee Tan

---

Conference Tutorials webpage: http://www.apsipa2015.org/ then click Tutorials tab

7th APSIPA ASC 2015

ASIA-PACIFIC SIGNAL AND INFORMATION PROCESSING ASSOCIATION ANNUAL SUMMIT AND CONFERENCE 2015

DECEMBER 16-19, 2015
HONG KONG

apsipa2015.org

Honorary General Chair
Wan-Chi Siu, Hong Kong Polytechnic University

General Co-Chairs
Kenneth Lam, Hong Kong Polytechnic University
Oscar Au, Hong Kong Univ. of Science & Tech.
Helen Meng, Chinese University of Hong Kong

Technical Program Co-Chairs
Changchun Bao, Beijing Univ. of Technology
Akira Hirabayashi, Ritsumeikan University
Jiwu Huang, Shenzhen University
Gwo Giun Lee, National Cheng Kung University
Daniel Lun, Hong Kong Polytechnic University
Tomoaki Ohtsuki, Keio University
Tomasz M. Rutkowski, University of Tsukuba
Sumei Sun, I2R, A*STAR

Finance Chair
Chris Chan, Hong Kong Polytechnic University

Secretary
Bonnie Law, Hong Kong Polytechnic University

Forum Co-Chairs
Wai-Kuen Cham, Chinese Univ. of Hong Kong
Homer Chen, National Taiwan University
King N. Ngan, Chinese University of Hong Kong
Ming-Ting Sun, University of Washington

Panel Session Co-Chairs
Shing-Chow Chan, University of Hong Kong
Dominic K.C. Ho, University of Missouri
Yo-Sung Ho, Gwangju Inst. of Science & Tech.
Yoshikazu Miyanaga, Hokkaido University

Special Session Co-Chairs
Mrityunjoy Chakraborty, India Inst. of Technology
Yui-Lam Chan, Hong Kong Polytechnic University
Lap-Pui Chau, Nanyang Technological University
Hsueh-Ming Hang, National Chiao-Tung University
Hitoshi Kiya, Tokyo Metropolitan University

Tutorial Co-Chairs
Waleed Abdulla, University of Auckland
Woon-Seng Gan, Nanyang Technological Univ.
Wing-Kuen Ling, Guangdong Univ. of Tech.
Lee Tan, Chinese University of Hong Kong

Registration Co-Chairs
Man-Wai Mak, Hong Kong Polytechnic University
Lai-Man Po, City University of Hong Kong

Publication Chair
Zheru Chi, Hong Kong Polytechnic University

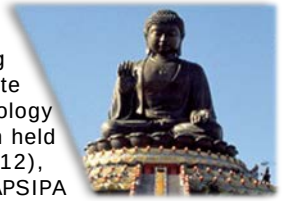
Publicity Co-Chairs
Kiyoharu Aizawa, University of Tokyo
Yui-Lam Chan, Hong Kong Polytechnic Univ.
Hing Cheung So, City University of Hong Kong
Mark Liao, IIS, Academia Sinica
Thomas Fang Zheng, Tsinghua University

Local Arrangement Co-Chairs
Edward Cheung, Hong Kong Polytechnic Univ.
Frank Leung, Hong Kong Polytechnic University

Advisory Committee
Chairs:
Sadaoki Furui, Toyota Technological Institute at Chicago
Wen Gao, Peking University
C.-C. Jay Kuo, University of Southern California
Haizhou Li, Inst. for Infocomm Research, A*STAR
Ray Liu, University of Maryland

Members:
Thierry Blu, Chinese University of Hong Kong
Shih-Fu Chang, Columbia University
Liang-Gee Chen, National Taiwan University
Li Deng, Microsoft Research
Takeshi Ikenaga, Waseda University
Kebin Jia, Beijing Univ. of Technology
Anthony Kuh, University of Hawaii
Antonio Ortega, University of Southern California
Soo-Chang Pei, National Taiwan University
Susanto Rahardja, National Univ. of Singapore
Yodchanan Wongsawat, Mahidol University
Chung-Hsien Wu, National Cheng Kung Univ.

CALL FOR PAPERS



APSIPA ASC 2015 will be the seventh annual conference organized by the Asia-Pacific Signal and Information Processing Association (APSIPA). Founded in 2009, APSIPA aims to promote research and education in signal processing, information technology and communications. Annual conferences have previously been held in Japan (2009), Singapore (2010), China (2011), the USA (2012), Taiwan (2013), and Cambodia (2014). The field of interest of APSIPA concerns all aspects of signal and information including processing, recognition, classification, communications, networking, computing, system design, security, implementation, and technology with applications to scientific, engineering and social areas. **Accepted papers in regular sessions and special sessions will be published in APSIPA ASC 2015 proceedings which will be submitted for inclusion into IEEE Xplorer as well as other Abstracting and Indexing (A&I) databases.**

The technical program includes the following tracks, but are not limited to:

1. Biomedical Signal Processing and Systems (BioSiPS)
Biomedical Signal and Information: Theory and Methods, Medical Information and Telecare Systems, Neural Systems and Applications, Bio-inspired Signal Processing and System, Biomedical Circuits and Systems.

2. Signal Processing Systems: Design and Implementation (SPS)
Nanoelectronics and Gigascale Systems, VLSI Systems and Applications, Embedded Systems, 3D Video Processing and Coding, High Efficiency Video Coding.

3. Image, Video and Multimedia (IVM)
Image/Video Processing, Coding and Analysis, Image/Video Storage, Retrieval and Authentication, Computer Vision and Graphics, Multimedia Systems and Applications.

4. Speech, Language and Audio (SLA)
Audio Processing, Speech Information Processing: Recognition, Synthesis, Understanding and Translation, Natural Language Processing: Language Modeling, Natural Language Generation/Understanding, Machine Translation.

5. Signal and Information Processing Theory and Methods (SIPTM)
Modern Signal Processing Theory and Method, Detection and Parameter Estimation, Array Processing and Multi-channels, Signal and Information Processing in Applications.

6. Wireless Communications and Networking (WCN)
Information and Network Security, Wireless Communications and Networking, Standards and Emerging Technology, RF and Antennas.

PAPER SUBMISSION

Prospective authors are invited to submit either full papers, up to 10 pages in length, or short papers up to 4 pages in length, where full papers will be for the single-track oral presentation and short papers will be mostly for poster presentation.

IMPORTANT DATES

Submission of Proposals for Special Session, Forum, Panel & Tutorial Sessions:	May 8, 2015
Submission of Full and Short Papers:	June 8, 2015
Submission of Papers in Special Sessions:	June 8, 2015
Notification of Papers Acceptance:	September 1, 2015
Submission of Camera-Ready Papers:	October 1, 2015
Author Registration Deadline:	October 1, 2015
Tutorial Session Date:	December 16, 2015
Summit and Conference Dates:	December 17-19, 2015

IEEE Signal Processing Society

Organizer/Sponsor: The Hong Kong Polytechnic University

Sponsor: Asia-Pacific Signal and Information Association (APSIPA)

Technical Co-Sponsor: IEEE Signal processing Society
IEEE Hong Kong Chapter of Signal Processing

Email: secretary@apsipa2015.org

APSIPA Who's Who

President: **Haizhou Li**, Institute for Infocomm Research, A*STAR, Singapore

President-Elect: **Wan-Chi Siu**, The Hong Kong Polytechnic University, Hong Kong

Past President: **C.-C. Jay Kuo,** University of Southern California, USA

VP - Conferences: **Susanto Rahardja**, National University of Singapore, Singapore

VP - Industrial Relations and Development: **Haohong Wang**, TCL Research America, USA

VP - Institutional Relations and Education Program:

Yo-Sung Ho, Gwangju Institute of Science and Technology, Korea

VP - Member Relations and Development: **Kenneth Lam,** The Hong Kong Polytechnic University, Hong Kong

VP - Publications: **Tatsuya Kawahara**, Kyoto University, Japan

VP - Technical Activities: **Oscar C. Au**, Hong Kong University of Science and technology, Hong Kong

Members-at-Large:

Waleed H. Abdulla, The University of Auckland, New Zealand

Kiyoharu Aizawa, The University of Tokyo, Japan

Mrityunjoy Chakraborty, Indian Institute of Technology, India

Kosin Chamnongthai, King Mongkut's University of Technology Thonburi, Thailand

Homer Chen, National Taiwan University, Taiwan

Woon-Seng Gan, Nanyang Technological University, Singapore

Wen Gao, Peking University, China

Hsueh-Ming Hang, National Chiao-Tung University, Taiwan

Anthony Kuh, University of Hawaii at Manoa, USA

K.J. Ray Liu, University of Maryland, USA

Tomoaki Ohtsuki, Keio University, Japan

Ming-Ting Sun, University of Washington, USA

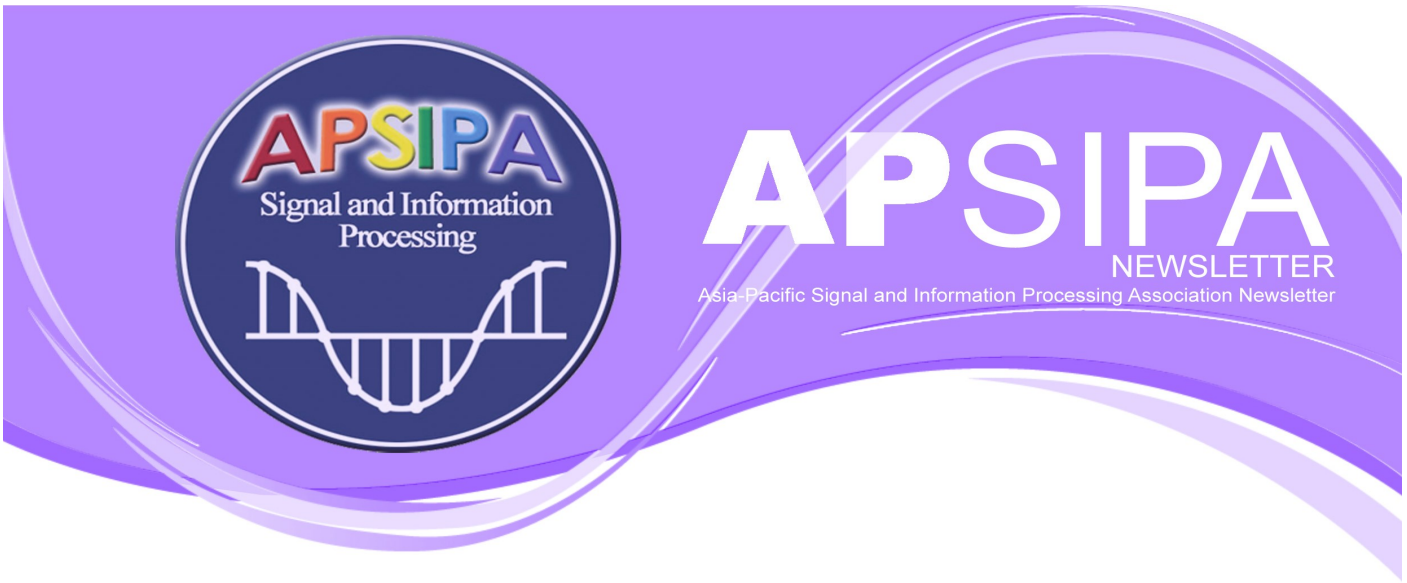
Headquarters

Address:

Asia Pacific Signal and Information Processing Association,
Centre for Signal Processing,
Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University,
Hung Hom, Kowloon, Hong Kong.

Officers:

Director: **Wan-Chi Siu**, email: enwcsiu@polyu.edu.hk
Manager: **Kin-Man Lam, Kenneth,**
email: enkmlam@polyu.edu.hk
Secretary: **Ngai-Fong Law, Bonnie,**
email: ennflaw@polyu.edu.hk
Treasurer: **Yuk-Hee Chan, Chris**,
email: enyhchan@polyu.edu.hk



APSIPA Newsletter Editorial Board Members

Waleed H. Abdulla (Editor-in-Chief), The University of Auckland, New Zealand.

Iman T. Ardekani, Unitec Institute of Technology, New Zealand.

Oscar C. Au, Hong Kong University of Science and technology, Hong Kong

Takeshi Ikenaga, Waseda University, Japan

Yoshinobu Kajikawa, Kansai University, Japan

Hitoshi Kiya, Tokyo Metropolitan University, Japan

Anthony Kuh, University of Hawaii, USA

Kenneth Lam, The Hong Kong Polytechnic University, Hong Kong

Bonnie Law (Deputy EiC), The Hong Kong Polytechnic University, Hong Kong

Haizhou Li, Institute for Infocomm Research, A*STAR, Singapore

Tomoaki Ohtsuki, Keio University, Japan

Woon-Seng Gan, Nanyang Technological University, Singapore

Yodchanan Wongsawat, Mahidol University, Thailand

Chung-Hsien Wu, National Cheng Kung University, Taiwan

Thomas Zheng, Tsinghua University, China

Are you an APSIPA member?

If not, then register online at

http://www.apsipa.org