Teknik Normalisasi Fitur Secara Adaptif untuk Sistem Pengenalan Ucapan Tahan Terhadap Gema

Adaptive Feature Normalization for Speech Recognition Robust Against Reverberation

Hilman Ferdinandus Pardede[1]

[1] *Pusat Penelitian Informatika, Lembaga Ilmu Pengetahuan Indonesia, Bandung, Indonesia*
Email: hilm001@lipi.go.id

Abstract

Reverberation degrades the performance of speech recognition significantly. Normalizing the features is arguably the most popular method to reduce the effect of reverberation on speech recognition. In previous works, q-log spectral mean normalization (q-LSMN) has been shown effective to remove convolutional and additive distortions in speech recognition. This method is an extension of traditional mean normalization on q-log domain. This method is extended in order to deal with reverberation and an adaptive technique to determine a suitable q in q-LSMN is proposed. Recognition results on digit recognition tasks for real recordings show the proposed method improves the robustness of speech recognition against reverberation. It is better than traditional techniques such as cepstral or log spectral mean normalization.

Keywords: *feature normalization, q-logarithm, reverberation, automatic speech recognition*

Abstrak

Gema menurunkan performa sistem pengenalan ucapan (SPU) atau *automatic speech recognition* secara signifikan. Salah satu teknik yang paling populer untuk mengurangi efek gema adalah dengan menormalisasi fitur pada SPU. Pada penelitian sebelumnya, q-log spectral mean normalization (q-LSMN) telah diperkenalkan untuk mengurangi efek distorsi aditif dan convolutif. Metode ini merupakan pengembangan teknik normalisasi konvensional pada domain q-log. Metode ini dikembangkan untuk mengurangi efek gema dan teknik adaptif untuk menentukan nilai q terbaik untuk q-LSMN diperkenalkan. Hasil percobaan pada pengenalan angka (*digit recognition*) menunjukkan bahwa teknik tersebut meningkatkan ketahanan SPU terhadap gema. Metode ini lebih baik dibandingkan metode normalisasi konvensional seperti *cepstral mean normalization* dan *log spectral mean normalization*.

Kata kunci: normalisasi fitur, q-logarithma, gema, sistem pengenalan ucapan

1. Pendahuluan

Aplikasi sistem pengenalan ucapan (SPU) atau *automatic speech recognition* pada kondisi bebas genggam (*hands-free*) yang memungkinkan manusia berinteraksi dengan SPU tanpa menggunakan tangan, semakin diminati belakangan ini [1]. Contohnya pada applikasi rumah otomatis, mobil pintar, atau robot. Pada aplikasi ini, SPU harus dapat beroperasi dimana terdapat gema/gaung, derau, ataupun adanya pembicara lain. Gema merupakan salah satu tantangan tersulit dan memiliki efek yang sangat besar dalam menurunkan performa SPU [2], [3], [4].

Received: 8 August 2016; Revised: 21 December 2016; Accepted: 23 September 2016; Published online: 22 February 2017
INKOM 2016/16-NO475 DOI:http://dx.doi.org/10.14203/j.inkom.475

Untuk meningkatkan performa SPU terhadap gema, berbagai metode telah diperkenalkan dalam beberapa dekade terakhir. Metode-metode ini secara umum dapat diklasifikasikan menjadi dua: metode berbasis *front-end* (FE) dan berbasis *back-end* (BE). Metode berbasis FE bertujuan untuk menghilangkan gema dari sinyal ucapan sehingga fitur yang dihasilkan mendekati fitur yang dihasilnya sinyal ucapan bersih (tanpa gema). Metode ini biasanya beroperasi di domain sinyal, spektrum, ataupun fitur. Di domain sinyal, contohnya adalah penggunaan *multi-microphone* [5], [6], [7] dan menemukan *inverse* dari *room impulse response* (RIR) [8], [9], [10]. Di domain spektrum, teknik speech enhancement seperti spectral subtraction, dan Ephraim Malah (EM) [11] biasanya digunakan. Sementara itu, Vector Taylor series (VTS) [12] and missing data theory [13] adalah beberapa contoh penggunaan

metode berbasis FE di domain fitur. Metode-metode berbasis FE biasanya memerlukan estimasi RIR, *reverberation time* (T60) dan/atau estimasi spektrum gema. Parameter-parameter ini umumnya bersifat non-stationary dan dapat berubah-ubah dipengaruhi berbagai faktor, misalnya pergerakan pembicara, variabilitas ruangan, dan karakteristik sinyal ucapan yang bersifat quasi-stationary. Oleh karena itu, sulit mendapatkan estimasi yang baik dari paramenter-parameter tersebut. Sementara itu, metode berbasis BE berusaha mengadaptasi model akustik yang biasanya diperoleh untuk keadaan tanpa gema, kepada kondisi dengan gema. Metode-metode berbasis BE antara lain dengan melatih kembali (*retraining*) model akustik dengan sinyal ucapan bergema [14], menggunakan *maximum a posteriori* (MAP) [15], dan *maximum likelihood linear regression* (MLLR) [16] sebagai teknik adaptasi. Akan tetapi, mengadaptasi model kedalam kondisi bergema mengakibatkan model akustik yang diperoleh menjadi sangat dipengaruhi oleh ruangan, sehingga performa SPU dapat menurun ketika SPU digunakan diruangan lain dengan parameter ruangan yang berbeda.

Mel-Frequency Cepstral Coeffient (MFCC) [17] dapat dikatakan sebagai fitur yang paling sering digunakan untuk SPU. MFCC diperoleh dengan melakukan analisis waktu pendek (sekitar 20-50 ms), dengan asumsi sinyal ucapan bersifat *stationary* selama periode tersebut. MFCC menggunakan skala Mel dan fungsi logaritma (log) untuk mengadopsi sistem pendengaran manusia dan mengompresi fitur. Fungsi log digunakan karena fungsi ini bersifat *homomorphic*, dapat mentransformasi operasi perkalian menjadi penjumlahan. MFCC menunjukkan performa yang baik untuk kondisi bersih, namun performanya sangat tidak tahan ketika terdapat gema, ataupun derau. Salah satu alasannya adalah penggunaan log yang sangat sensitif pada daerah dimana spektrum memiliki energi yang rendah. Daerah ini merupakan tempat dimana informasi dari sinyal ucapan berada. Selain itu, pola sinyal ucapan bersifat kompleks. Unit dari ucapan berupa kata atau fonem memiliki durasi yang berbeda-beda yang dapat lebih pendek atau lebih panjang dari periode analisis pada MFCC (yakni sekitar 20ms). Oleh karena itu, pada domain spektral, komponen spektral dari sinyal ucapan memiliki korelasi satu sama lain. Ketika sinyal ucapan dipengaruhi gema, energi spektra menjadi tersebar lebih panjang, mengakibatkan korelasinya menjadi lebih tinggi.

Keterbatasan MFCC menyebabkan banyak penelitian telah memperkenalkan fitur-fitur lain sebagai pengganti MFCC, misalnya minimum variance distortion-less response (MVDR) [18], [19]. Fitur ini diperoleh dengan menggunakan FIR filter pada sinyal ucapan sehingga sinyal keluaran memiliki unit gain. Dengan demikian, pengaruh bias dan variance dapat dikurangi. Frequency-domain linear prediction (FDLP) adalah contoh fitur lain [20]. Pada fitur ini linear prediction digunakan pada sinyal pita sempit (narrow-band) untuk memperoleh temporal-envelope sinyal ucapan tersebut. Sehingga dengan menormalisasi sinyal tersebut, efek gema dapat dikurangi. Sinyal pita sempit digunakan agar periode analisis lebih panjang dibandingkan T60 sehingga gema dapat diasumsikan sebagai gain terhadap sinyal bersih.

Pendekatan lain untuk fitur baru adalah penggunaan fungsi akar menggantikan log pada ekstraksi fitur [21]. Perceptually linear prediction (PLP) [22] adalah salah satu contoh penggunaan fungsi akar. Power normalized cepstral coefficient (PNCC) [23] adalah contoh lain fitur yang menggunakan fungsi akar. PNCC telah terbukti lebih tahan terhadap distorsi dari lingkungan dibandingkan MFCC, PLP, dan MVDR, serta teknik speech enhancement seperti VTS [24], [25]. PNCC memiliki proses ekstraksi fitur yang menyerupai MFCC, kecuali dalam tiga hal. Pertama, PNCC menggunakan Gammatone filterbank sedangkan MFCC menggunakan mel filterbank. Kedua, PNCC memiliki teknik penghilang noise: medium duration power bias subtraction and power peak normalization dan ketiga, PNCC menggunakan fungsi akar menggantikan log pada MFCC.

Efektifitas fitur berbasis fungsi akar dibandingkan log dikarenakan fungsi akar lebih tidak sensitif terhadap perubahan pada spektra berenergi rendah seperti log [26]. Akan tetapi performa fitur berbasis fungsi akar sangat tergantung dengan teknik normalisasi yang digunakan [27], [28]. Teknik normalisasi konvensional [27], bukanlah teknik terbaik untuk fitur ini karena sifat properti fungsi akar tidak sama dengan log.

Fungsi q-logarithma (q-log) adalah merupakan contoh fungsi akar dan generalisasi fungsi natural logaritma (log). Fungsi ini banyak digunakan dalam statistika Tsallis [29], [30]. Dalam statistika Tsallis, fungsi ini digunakan untuk menjelaskan fenomena non-extensive pada sistem kompleks. Teknik normalisasi berbasi fungsi q-log based telah diperkenalkan [31], [32]. Teknik ini dinamakan q-log spectral mean normalization (q-LSMN), menggunakan sifat/properti fungsi q-log dan terbukti efektif mengurangi efek derau additif dan convolutif. Pada studi ini, hanya satu nilai q yang digunakan yang ditentukan secara empiris. Beberapa studi mengindikasikan, penggunaan akar lebih dari satu lebih baik dibandingkan penggunaan nilai akar tunggal [33], [34].

Pada makalah ini, metode q-LSMN dikembangkan untuk mengatasi masalah gema pada SPU dan teknik adaptif untuk menentukan nilai q diperkenalkan. Teknik adaptif ini memungkinkan

dilakukan berbagai kompresi pada bagian berbeda sinyal ucapan. Hasil evaluasi menunjukkan teknik ini lebih baik dibandingkan penggunaan nilai q tunggal dan berbagai teknik normalisasi lainnya.

2. Formulasi Masalah: Efek Gema Terhadap Sinyal Ucapan

Gema, yang ditandai dengan room impulse response (RIR), biasanya dimodelkan memiliki relasi konvolusi dengan sinyal ucapan pada domain waktu. Hubungan antara sinyal bersih x(t), terdistorsi oleh gema dengan RIR h(t) dan sinyal terdistorsi y(t) dapat diformulasikan sebagai berikut:

$$y(t) = h(t) * x(t). \quad (1)$$

RIR ditentukan oleh parameter *reverberation time* (T60). T60 adalah waktu yang dibutuhkan oleh sinyal ucapan untuk berkurang sebesar 60 desibel dibawah level normalnya. Walaupun sinyal ucapan dan gema adalah konvolutif di domain waktu, hubungan keduanya tidak menjadi multiplikatif di domain frekuensi ketika T60 lebih besar dibandingkan periode analisisnya yang biasanya sebsar 2050 ms. Untuk memudahkan analisis, RIR dapat dibagi menjadi 2, gema awal (*early reverberation*) dan gema akhir (*late reverberation*) berdasarkan kapan sinyal tersebut mencapai microphone. Secara matematis dapat dituliskan sebagai berikut:

$$h(t) = \begin{cases} h_e(t) & \text{for } 0 \leq t < t_d; \\ h_l(t) & \text{for } t \geq t_d, \end{cases} \quad (2)$$

dimana h_e dan h_l merepresentasikan gema awal dan gema akhir, t_d adalah nilai yang digunakan untuk membedakan gema awal dan gema akhir. Persamaan 1 dapat ditulis menjadi:

$$y(t) = \sum_{\tau=0}^{t_d} h_e(\tau)s(t-\tau) + \sum_{\tau=t_d}^{T} h_l(\tau)s(t-\tau). \quad (3)$$

Dengan menggunakan Short-time Fourier Transform (STFT) terhadap Persamaan 3, dan dengan mengasumsikan t_d lebih kecil dari panjang jendela analisis, Persamaan 3 dapat dituliskan dalam domain frekuensi sebagai berikut:

$$|Y(m,k)| = |X(m,k)||H_e(k)| + \lambda(m,k), \quad (4)$$

dimana $|Y(m,k)|$ dan $|X(m,k)|$ adalah spektra magnitude dari $y(t)$ dan $x(t)$ pada frame ke-m dan index frekuensi k. $|H_e(k)|$ adalah spectra magnitide untuk $h_e(t)$ dan notasi $\lambda(m,k)$ merepresentasikan spektra gema akhir yang dapat dijabarkan sebagai berikut:

$$\lambda(m,k) = \sum_{t=t_d}^{L} |X(m-t,k)||H_l(m,k)|, \quad (5)$$

dimana $|H_l(m,k)|$ adalah spektra magnitude dari $h_l(t)$. Berdasarkan persamaan 4, gema dapat memiliki relasi multiplikatif dan additif terhadap sinyal ucapan ketika RIR memiliki T60 yang panjang. Gema awal bersifat stationary dan memiliki pengaruh kecil terhadap performa SPU sedangkan gema akhir bersifat non-stationary dan memiliki pengaruh yang signifikan terhadap penurunan performa SPU. Untuk memudahkan solusi untuk persamaan 4, $\lambda(m,k)$ dan $|X(m,k)|$ diasumsikan tidak tercorelasi.

3. Normalisasi Fitur pada Sinyal Ucapan Terdistorsi Gema

3.1. Metode Normalisasi Fitur Konvensional

Metode normalisasi konvensional dilakukan dengan mengurangkan fitur dengan nilai rata-rata fitur tersebut. Metode normalisasi tradisional seperti Contoh teknik normalisasi konvensional adalah Cepstral mean normalsation (CMN) [35], [36] dan log spectral mean normalisation (LSMN) [37]. Kedua metode ini bekerja erdasarkan prinsip yang sama namun dilakukan di domain yang berbeda yaitu cepstral dan log. Metode ini efektif untuk menghilangkan efek gema awal, yaitu hanya ketika T60 adalah relatif singkat [38], [39], [40]. Namun, metode ini tidak efektif menghilangkan gema akhir seperti dijelaskan sebagai berikut. Sinyal ucapan dan gema adalah multiplikatif di domain spektral sehingga keduanya menjadi aditif di domain log. Dinotasikan:

$$\alpha(m,k) = 1 + \frac{\lambda(m,k)}{|X(m,k)||H_e(k)|}. \quad (6)$$

Maka persamaan 4 menjadi:

$$|Y(m,k)| = |X(m,k)||H_e(k)|\alpha(m,k). \quad (7)$$

Dengan mengambil log dari persamaan 7, maka 7 dapat dituliskan menjadi:

$$\mathbf{y}(m,k) = \mathbf{x}(m,k) + \mathbf{h}_e(k) + \boldsymbol{\alpha}(m,k), \quad (8)$$

dimana $\mathbf{y}$, $\mathbf{x}$, $\mathbf{h}_e$, dan $\boldsymbol{\alpha}$ adalah log spektrum dari Y, X, H_e dan α. Dengan menormalisasi persamaan 8, diperoleh:

$$\tilde{\mathbf{y}}(m,k) = \tilde{\mathbf{x}}(m,k) + \tilde{\boldsymbol{\alpha}}(m,k) \quad (9)$$

dimana $\tilde{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}$ dan $\tilde{\boldsymbol{\alpha}} = \boldsymbol{\alpha} - \overline{\boldsymbol{\alpha}}$, adalah hasil normalisasi *mean* dari $\mathbf{x}$ and $\boldsymbol{\alpha}$. Karena gema awal bersifat stationary, melakukan normalisasi dapat menghilangkannya. Akan tetapi, gema akhir tidak dapat dihilangkan.

3.2. Q-Logaritma

Fungsi q-log dari variable x didefinisikan sebagai berikut:

$$\log_q(x) = \frac{x^{1-q} - 1}{1-q}. \quad (10)$$

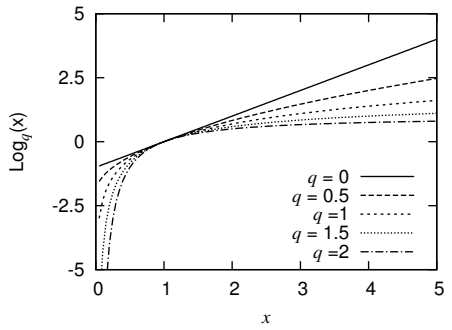

Figure 1. Fungsi q-logarithma terhadap variable x untuk berbagai nilai q.

Fungsi ini asimtotik mendekati logaritma natural ketika q mendekati 1 seperti terlihat pada Gambar 1. Inverse dari fungsi ini disebut fungsi q-eksponensial, didefinisikan sebagai berikut:

$$\exp_q(x) = (1 + (1-q)x)^{\frac{1}{1-q}}. \tag{11}$$

Fungsi q-log telah diterapkan di beberapa bidang, misalnya dalam pengolahan sinyal ucapan [41], dalam statistika Tsallis [29], [30]. Fungsi q-log menyediakan platform non-aditif saat $q \neq 1$ [42], [43]. Dalam Tsallis statistik, platform ini digunakan untuk menjelaskan fenomena non-ekstensif yang banyak ditemukan dalam berbagai sistem yang kompleks dalam fisika, biologi, ekonomi, keuangan, dll. Fenomena non-ekstensif ini disebabkan adanya korelasi yang belum diketahui. Parameter q digunakan dan ditentukan secara empiris agar perilaku sistem dapat dijelaskan.

3.3. Efek Normalisasi Konvensional Terhadap Fitur Berbasis Q-Log

Untuk fitur berbasis q-log, Hasi normalisasi fitur dengan teknik normalisasi konvensional menghasilkan fitur sebagai berikut, dengan asumsi sinyal ucapan dan gema akhir tidak berkorelasi (indeks frame dan frekuensi diabaikan):

$$\tilde{\mathbf{y}}_q = \left(1 + (1-q)\mathbf{h}_{e_q}\right)\left(\tilde{\mathbf{x}}_q + \tilde{\boldsymbol{\alpha}}_q + (1-q)\tilde{\mathbf{x}}_q\tilde{\boldsymbol{\alpha}}_q\right), \tag{12}$$

dimana $\mathbf{h}_{e_q}$ adalah spektrum q-log dari $|H_e|$, $\tilde{\mathbf{x}}_q$ dan $\tilde{\boldsymbol{\alpha}}_q$ adalah fitur yang ternormalisasi dari $\mathbf{x}_q$ dan $\boldsymbol{\alpha}_q$.Dari persamaan 12 jelas bahwa menerapkan metode normalisasi konvensional tidak menghilangkan komponen gema awal dan akhir. Gema awal masih menjadi gain terhadap sinyal ucapan. Oleh karena itu normalisasi gain biasanya diterapkan pada fitur berbasis fungsi root seperti pada PNCC [23]. Pada PNCC, Dalam PNCC, power peak normalisation dan power bias subtraction diterapkan sebelum dinormalisasi. Ini menyebabkan efek derau dan derau konvolutif seperti gema awal berkurang.

3.4. Q-Log Mean Normalization (Q-LSMN)

Q-LSMN pada $\mathbf{y}_q$ diformulasikan sebagai berikut:

$$\breve{\mathbf{y}}_q = \frac{\mathbf{y}_q - \overline{\mathbf{y}}_q}{1 + (1-q)\overline{\mathbf{y}}_q}, \tag{13}$$

dimana $\breve{\mathbf{y}}_q$ adalah hasil normalisasi setelah q-LSMN dan $\overline{\mathbf{y}}_q$ adalah nilai rata-rata (*mean*) dari $\mathbf{y}_q$. Perlu diingat, akibat perbedaan sifat dan properti dari q-log dan log maka rata-rata aritmatikal $\overline{\mathbf{y}}_q$ di domain spektrum bukanlah rata-rata geometris ketika $q \neq 1$, melainkan diformulasikan sebagai berikut:

$$\begin{aligned}\overline{\mathbf{y}}_q &= \frac{1}{M}\sum_{m=1}^{M}\mathbf{y}_q(m,k) \\ &= \frac{1}{M}\log_q\left(|X(1,k)| \times_q ... \times_q |X(M,k)|\right),\end{aligned} \tag{14}$$

dimana M adalah jumlah total dari frame dan $\times_q$ merupakan generalisasi dari operator perkalian [43], [42] yang didefinisikan sebagai berikut:

$$a \times_q b = \left(a^{1-q} + b^{1-q} - 1\right)^{\frac{1}{1-q}} \tag{15}$$

Dari persamaan 14 dapat dilihat bahwa nilai rata-rata $\mathbf{y}$ berada diantara nilai rata-rata aritmatis dan rata-rata geometris ketika q adalah antara 0 dan 1.

Ketika q-LSMN diterapkan kepada sinyal ucapan terdistorsi gema, dengan mengambil q-log pada persamaan 7 dan menerapkan q-LSMN seperti pada persamaan 13, dan mengasumsikan sinyal ucapan dan gema tidak terkorelasi, maka, spektrum ternomalisasi hasil q-LSMN adalah sebagai berikut:

$$\begin{aligned}\breve{\mathbf{y}}_q =& \frac{\breve{\mathbf{x}}_q}{1 + (1-q)\overline{\boldsymbol{\alpha}}_q} \\ &+ \frac{\tilde{\boldsymbol{\alpha}}_q + (1-q)\tilde{\mathbf{x}}_q\tilde{\boldsymbol{\alpha}}_q}{\left(1 + (1-q)\overline{\boldsymbol{\alpha}}_q\right)\left(1 + (1-q)\overline{\mathbf{x}}_q\right)}.\end{aligned} \tag{16}$$

Dapat dilihat bahwa hasil normalisasi menggunakan q-LSMN tidak memerlukan normalisasi *gain* seperti pada PNCC. Berdasarkan persamaan (16), maka ketika $q = 1$, q-LSMN identik dengan LSMN dan $\boldsymbol{\alpha}_q$ yang merupakan representasi dari gema akhir tidak dapat dihilangkan. Ini juga mengonfirmasi keterbatasan fungsi log untuk gema akhir. Ketika $q < 1$ digunakan, maka nilai rata-rata aritmatis dari $\mathbf{y}_q$ lebih tinggi dari nilai rata-rata $q = 1$ dan efek q-LSMN dapat dianalisa sebagai berikut. Jika sinyal ucapan memiliki energi yang lebih besar dibandingkan gema, atau secara matematis dapat dituliskan $\breve{\mathbf{x}}_q \gg \tilde{\boldsymbol{\alpha}}_q$, maka energi sinyal bersih dapat menekan gema tersebut (lihat bagian kedua persamaan (16). Akibatnya efek gema akhir akan berkurang. Ini bisa terjadi dibagian sinyal ucapan yang berisi suara (seperti bunyi vokal atau konsonan bersuara seperti nasal). Sebaliknya ketika

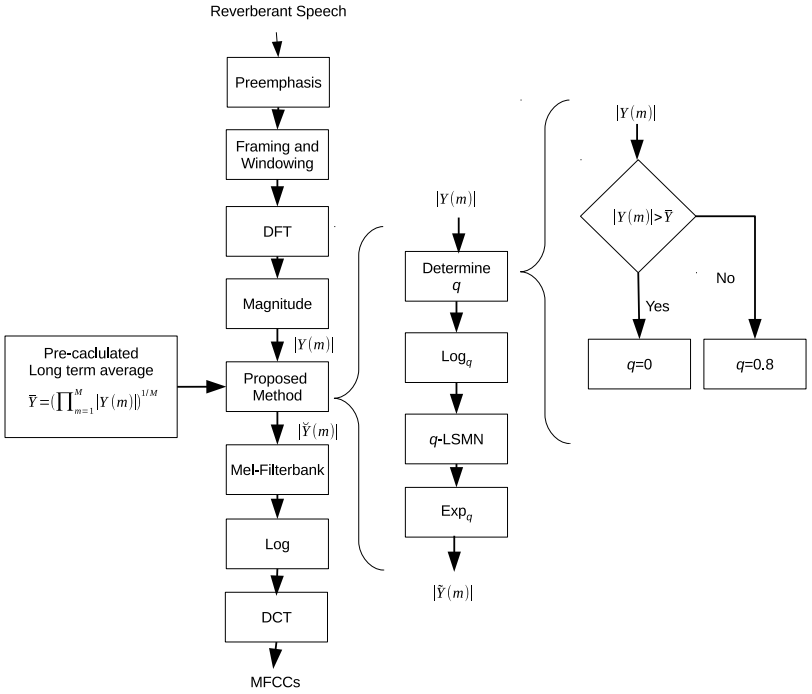

Figure 2. Diagram blok proses ekstraksi dari metode yang diusulkan.

energi sinyal ucapan lebih kecil, $\breve{\mathbf{x}}_q \ll \tilde{\boldsymbol{\alpha}}_q$, maka bagian pertama dari persamaan (16) akan menjadi sangat kecil dan sinyal menjadi didominasi oleh $\tilde{\boldsymbol{\alpha}}_q$ yang akan teramplifikasi. Ini umumnya terjadi pada sinyal yang berisi konsonan tidak bersuara seperti bunyi letup (b, p, t, d) dan desis (f, s). Hal ini mengindikasikan bahwa nilai q harus dipilih secara hati-hati, dan nilai q harus dibedakan untuk sinyal yang memiliki energi rendah dan energi tinggi.

Dimotivasi hal ini, maka diusulkan penggunaan dua nilai q yang berbeda untuk daerah berenergi tinggi dan rendah. Untuk mengidentifikasi hal tersebut dapat dilakukan dengan mengenali apakah sinyal berisi informasi konsonan atau vokal. Akan tetapi hal tersebut, sehingga untuk mengidentifikasi suatu sinyal berdasarkan apakah sinyal tersebut merupakan puncak dari spektrum atau lembah. Tentu hal ini dengan asumsi bahwa efek gema tidak mengubah informasi lembah atau puncak suatu sinyal. Lebih detail mengenai metode ini akan dijelaskan dibagian selanjutnya.

4. Metode

Pada bagian sebelumnya, telah ditunjukkan bahwa penerapan q yang berbeda untuk bagian yang berbeda dari sinyal ucapan berpotensi meningkatkan ketahanan fitur terhadap gema dibandingkan penggunaan nilai q tunggal. Hal ini menjadi motivasi untuk mengembangkan teknik adaptif untuk menentukan nilai q. Fokus makalah ini adalah penggunaan dua nilai q yang berbeda yang akan diterapkan pada lembah spektrum dan puncak spektrum. Penentuan apakan suatu spektrum masuk kategori lembah atau puncak adalah dengan menentukan apakah *power* dari spektrum dari spektrum tersebut lebih rendah atau lebih tinggi dari nilai rata-rata geometris dari setiap ucapan. Ketika power spektrum lebih rendah dari nilai rata-rata tersebut, spektrum tersebut dikategorikan lembah sedangkan ketika power spektrum lebih besar dikategorikan sebagai puncak. Nilai $q = 0,8$ diberikan untuk daerah lembah, sedangkan $q = 0$ untuk puncak spektrum. Nilai ini berdasarkan pengamatan empiris bahwa pasangan nilai ini mencapai kinerja terbaik. Metode ini dinamakan teknik adaptif q-log spectral mean normalization (dinotasikan sebagai q-LSMN_A).

Gambar 2 menunjukkan ekstraksi fitur dari q-LSMN_A. Pertama, sinyal ucapan dilewatkan melalui filter pre-emphasis, $1 - 0,97z^{-1}$ dan kemudian Hamming window diterapkan pada output dari filter pre-emphasis. Untuk setiap frame, panjang jendela (*frame-length*) adalah 25 ms dan pergeseran setiap frame (*frame-shift*) adalah 10 ms. Untuk setiap frame,, Fast Fourier Transform (FFT) diterapkan dan kemudian diambil spektrum power dengan mengkuadratkan magnitude dari keluaran FFT. Setelah itu, untuk setiap frame ditentukan apakah nilai spektrum power lebih tinggi atau lebih rendah dari rata-rata geometrisnya. Nilai rata-rata ini ditentukan secara offline untuk setiap kalimat. Kemudian nilai q yang sesuai dipilih. Kemudian,

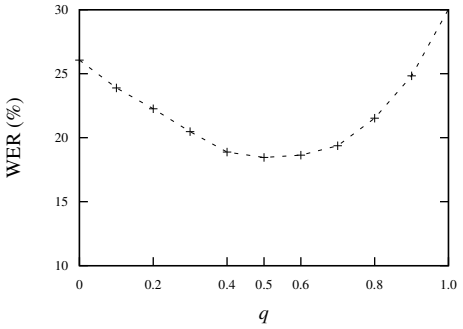

Figure 3. Kinerja q-LSMN ketika hanya satu nilai q yang diterapkan. Hasil yang ditampilkan adalah nilai rata-rata WER untuk semua mikrofon. Kinerja terbaik diperoleh ketika $q = 0,5$

spektrum ditransformasi ke domain q-log dan q-LSMN diterapkan sesuai dengan nilai q masing-masing spektrum. Setelah normalisasi, fitur yang kemudian diubah kembali ke domain spektral. Dalam makalah ini, MFCC tetap digunakan sebagai fitur akhir, untuk menunjukkan bahwa efek pada kinerja ASR dikarenakan oleh metode normalisasi, bukan karena efek transformasi ke domain q-log semata.

5. Percobaan

5.1. Pengaturan Percobaan

Untuk evaluasi metode ini, Corpus Aurora-5 digunakan [44]. Untuk pelatihan (*training*), digunakan standar set dari Aurora-5 untuk kondisi training bersih, berisi 8623 kalimat. Sedangkan untuk *testing* (Pengujian) digunakan set *Meeting Recorder Digit* (MRD). MRD adalah bagian dari Aurora-5 basis data yang terdiri dari rekaman nyata 2400 ucapan-ucapan dari 24 pembicara dalam ruang rapat, dengan total total 7800 kalimat. Setiap set direkam menggunakan 4 mikrofon yang berbeda (berlabel 6, 7, E, F) ditempatkan di tengah-tengah meja. Data ini mengandung sedikit derau.

5.2. Sistem Pengenalan Ucapan (SPU)

SPU dibangun berbasis Hidden Markov Model (HMM) menggunakan Hidden Markov Model toolkit (HTK) [45]. Sebagai fitur, digunakan 39 dimensi fitur yang meliputi 13 statis MFCC, (dengan *zeroth cepstra*) ditambah turunan pertama dan keduanya (fitur delta dan delta-delta). Kinerja ASR adalah diukur sebagai tingkat kesalahan kata (*word error rate* atau disingkat WER). Sebagai perbandingan, berbagai metode dan fitur lain digunakan, antara lain ceptral mean normalisation (CMN), mean variance normalisation (MVN), dan PNCC. Untuk PNCC, kami menerapkan dimensi yang sama (13 dimensi fitur statis + pertama dan turunan kedua).

Table 1. Perbandingan kinerja q-LSMN_A dan q-LSMN pada $q = 0,5$ (q-LSMN memiliki kinerja terbaik pada $q = 0,5$ ketika hanya satu nilai q yang dipakai)

Method	MIC				AVE
	6	7	E	F	
Q-LSMN_A	10,97	16,66	17,49	12,78	14,48
Q-LSMN	14,55	21,61	21,81	15,86	18,46

Table 2. Perbandingan kinerja q-LSMN_A dan beberapa metode-metode lain

Method	MIC				AVE
	6	7	E	F	
MFCC	34,15	51,70	36,87	31,49	38,55
CMN	25,33	35,85	33,36	25,48	30,01
MVN	25,74	33,71	31,93	26,50	29,47
PNCC	10,37	14,31	16,00	13,15	13,46
Q-LSMN_A	10,97	16,66	17,9	12,78	14,48

Untuk HMM, digunakan 8623 kalimat dengan rasio sampling (*sampling rate*) 8 kHz untuk melatih HMM untuk setiap digit. Setiap digit dimodelkan menggunakan HMM kiri-ke-kanan yang terdiri dari 16 *state* dan masing-masing state dimodelkan dengan Gaussian Mixture model (GMM) dengan 4 mixtures. Untuk model jeda, "sil", dimodelkan dengan tiga states HMM dan masing-masing state dimodelkan dengan GMM yang memiliki 4 *mixtures*.

5.3. Hasil Percobaan dan Analisa

Gambar 3 menunjukkan kinerja q-LSMN ketika hanya 1 nilai q yang diterapkan. Peningkatan performa dapat dicapai ketika $q \neq 1$, ketika q-log sama dengan log. Hasil ini mengkonfirmasi penggunaan fungsi power lebih baik dibandingkan dengan log. Dapat dilihat performa terbaik diperoleh ketika $q = 0,5$. Berdasarkan Tabel 5.3., q-LSMN_A mendapatkan pengurangan rasio WER (*error rate reduction* atau disingkat ERR) sebesar 51,80% dibandingkan q-LSMN ketika $q = 1$ dan 21,54% EER untuk $q = 0,5$, dimana q-LSMN mendapatkan performa terbaik.

Dibandingkan metode lainnya seperti CMN, LSMN, dan MVN, q-LSMN_A menghasilkan performa lebih baik dibandingkan ketiganya (Table 2). Hasil ini menunjukkan q-LSMN_A lebih efektif untuk mengurangi efek gema dibandingkan CMN dan LSMN. CMN dan LSMN (sama dengan q-LSMN ketika $q = 1$) memiliki performa yang sedikit berbeda. Hal ini tidak begitu mengherankan karena operasional CMN dan LSMN memiliki domain operasi yang berbeda: CMN beroperasi

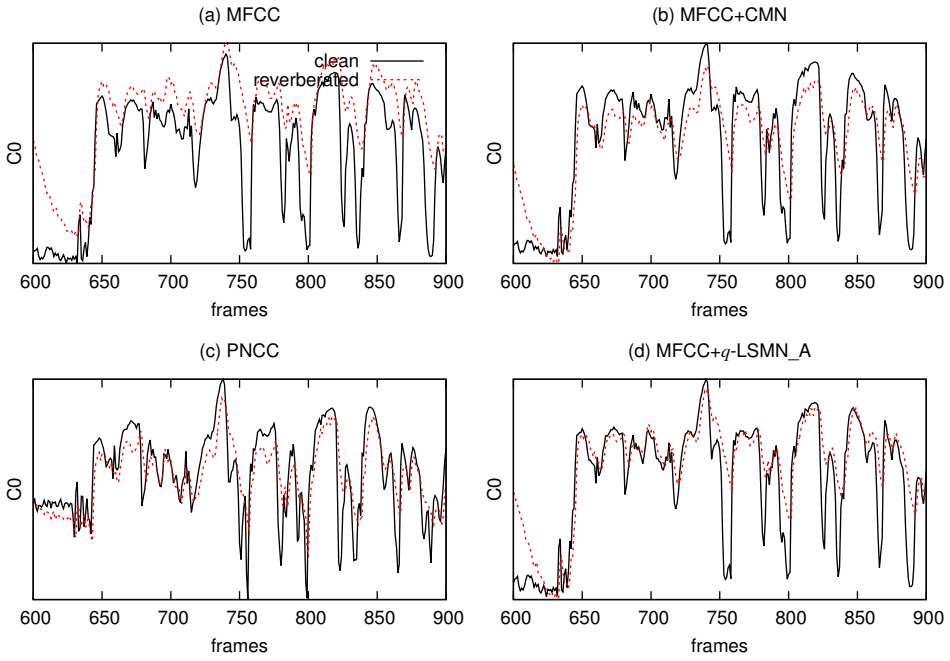

Figure 4. Perbadingan C0 (*zeroth cepstral*) dari sinyal ucapan bersih dan sinyal ucapan terkontaminasi gema.

pada mel filterbank sementara LSMN pada linear domain spektral, dan karenanya panjang dari kanal frekuensi yang berbeda. Dibandingkan dengan CMN, q-LSMN_A mencapai 51,75% ERR.

Secara rata-rata, q-LSMN_A sedikit lebih buruk dari PNCC. Hasil ini dapat dijelaskan menggunakan Gambar 4. Pada gambar ini, ketidakcocokan (*mismatch*) antara sinyal ucapan bersih dan bergema di *zeroth cepstra* (C0) untuk beberapa fitur: MFCC, MFCC + CMN, MFCC + q-LSMN_A, dan PNCC, dibandingkan. Terlihat jelas bahwa q-LSMN_A memiliki mismatch (ketidakcocokan) lebih kecil dari CMN. Dibandingkan dengan PNCC, dapat dilihat bahwa q-LSMN_A memiliki mismatch yang lebih kecil untuk puncak spektrum, tetapi mismatch untuk lembah spektral lebih besar. Namun demikian, hasil ini juga mengonfirmasi bahwa keuntungan menerapkan $q = 0$ pada puncak spektrum. Pada lembah spektral, q-LSMN_A terlihat tidak seefektif pada puncak spektrum. Ini mungkin dikarenakan energi dari frame sebelumnya dapat mempengaruhi seluruh ucapan karena ekor gema yang sangat panjang. Ini terjadi pada kondisi apabila SPU digunakan pada ruangan dengan ukuran yang besar. Oleh karena itu, energi dari lembah dapat bertambah dan mereka salah diidentifikasi sebagai puncak dan nilai q yang dipilih menjadi tidak tepat. Oleh karena itu, spektrum menjadi *overcompressed* dan mismatch akan menjadi lebih besar. Untuk PNCC, perlu diingat bahwa selain normalisasi, PNCC juga mencakup beberapa proses tambahan dalam proses ekstraksi fitur diantaranya teknik menghilangkan derau seperti power peak normalization dan power bias subtraction. Proses-proses tambahan ini ini juga bisa berkontribusi atas performa PNCC.

6. Kesimpulan

Dalam makalah ini, metode q-LSMN untuk mengurangi efek gema pada SPU telah diperkenalkan. Metode ini merupakan pengembangan dari q-LSMN yang merupakan teknik normalisasi dalam kerangka fungsi q-log. Tenik adaptif untuk menentukan nilai q merupakan kontribusi terutama pada makalah ini. Penggunaan dua nilai q: $q = 0$ ketika spektrum power lebih tinggi dari nilai rata-rata geometrinya sementara $q = 0,8$ untuk lembah spektrum dimana nilai spektrum power lebih rendah dibandingkan nilai rata-rata spektrum. Hasil eksperimen menunjukkan metode ini lebih ampuh untuk menghilangkan efek gema pada SPU dibandingkan ketika hanya satu nilai q yang digunakan. Metode ini juga lebih baik daripada teknik normalisasi tradisional seperti CMN dan MVN dan memiliki performa sebanding kinerja dengan PNCC.

Metode ini belum dapat diaplikasikan secara real-time karena membutuhkan informasi tentang spektrum di masa depan. Oleh karena itu teknik impelemtasi real-time menjadi rencana masa depan studi ini. Pengamatan empiris terhadap hasil percobaan mengindikasikan adanya pengaruh luas ruangan (hal ini berhubungan dengan parameter T60) dengan nilai q yang optimum. Hal ini menarik untuk diselidiki untuk penelitian di masa yang akan datang.

References

[1] M. Wölfel and J. McDonough, *Distant speech recognition*. Wiley, 2009.

[2] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 114–126, Nov 2012.

[3] A. Sehr, E. A. Habets, R. Maas, and W. Kellermann, "Towards a better understanding of the effect of reverberation on speech recognition performance," in *Internat. Workshop on Acoustic Echo and Noise Control*, 2010.

[4] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: An overview of challenge systems and outcomes," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, Dec 2013, pp. 162–167.

[5] J. Dennis and T. H. Dat, "Single and multi-channel approaches for distant speech recognition under noisy reverberant conditions: I2r's system description for the aspire challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Dec 2015, pp. 518–524.

[6] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170 – 180, 2014.

[7] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication*, vol. 25, no. 13, pp. 75 – 95, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639398000302

[8] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, no. 2, pp. 145–152, Feb 1988.

[9] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 5177–5181.

[10] S. Mosayyebpour, M. Esmaeili, and T. A. Gulliver, "Single-microphone early and late reverberation suppression in noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 2, pp. 322–335, Feb 2013.

[11] F. Xiong, B. T. Meyer, and S. Goetze, "A study on joint beamforming and spectral enhancement for robust speech recognition in reverberant environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 5043–5047.

[12] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, vol. 2, Atlanta, USA, May 1996, pp. 733 –736.

[13] K. J. Palomaki, G. J. Brown, and J. Barker, "Missing data speech recognition in reverberant conditions," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, May 2002, pp. I–65–I–68.

[14] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Training of hmm with filtered speech material for hands-free recognition," in *IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, Mar 1999, pp. 449–452 vol.1.

[15] J. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr 1994.

[16] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[17] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357 – 366, aug 1980.

[18] S. Dharanipragada and B. Rao, "Mvdr based feature extraction for robust speech recognition," in *IEEE Internat. Conf. on Acoustics, Speech and Signal Processing*, vol. 1, 2001, pp. 309–312 vol.1.

[19] M. Wolfel and J. McDonough, "Minimum variance distortionless response spectral estimation," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept 2005.

[20] S. Thomas, S. Ganapathy, and H. Hermansky, "Recognition of reverberant speech using frequency domain linear prediction," *IEEE Signal Processing Letters*, vol. 15, pp. 681–684, 2008.

[21] J. Lim, "Spectral root homomorphic deconvolution system," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 3, pp. 223 – 233, jun 1979.

[22] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[23] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 4101–4104.

[24] F. Kelly and N. Harte, "Auditory features revisited for robust speech recognition," in *Internat. Conf. Pattern Recognition*, Aug 2010, pp. 4456–4459.

[25] G. Sarosi, M. Mozsary, P. Mihajlik, and T. Fegyo, "Comparison of feature extraction methods for speech recognition in noise-free and in traffic noise

environment," in *Conf. Speech Technology and Human-Computer Dialogue*, May 2011, pp. 1–8.

[26] P. Alexandre and P. Lockwood, "Root cepstral analysis: A unified view. application to speech processing in car noise environments," *Speech Commun.*, vol. 12, no. 3, pp. 277 – 288, 1993.

[27] S. Baek and H. Kang, "Mean normalization of power function based cepstral coefficients for robust speech recognition in noisy environment," in *IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, May 2014, pp. 1735–1739.

[28] M. J. Hunt, "Spectral signal processing for ASR," in *IEEE Workshop Automatic Speech Recognition and Understanding*, Colorado, USA, 1999, pp. 17–25.

[29] C. Tsallis, "Possible generalization of boltzmann-gibbs statistics," *J. Stat. Phys.*, vol. 52, pp. 479–487, 1988.

[30] ——, "Nonadditive entropy: The concept and its use," *Eur. Phys. J. A.*, vol. 40, pp. 257–266, 2009, 10.1140/epja/i2009-10799-0. [Online]. Available: http://dx.doi.org/10.1140/epja/i2009-10799-0

[31] H. Pardede, K. Iwano, and K. Shinoda, "Feature normalization based on non-extensive statistics for speech recognition," *Speech Commun.*, vol. 55, no. 5, pp. 587 – 599, 2013.

[32] H. F. Pardede and K. Shinoda, "Generalized-log spectral mean normalization for speech recognition," in *Interspeech*, 2011, pp. 1645–1648.

[33] P. Lockwood and P. Alexandre, "Root adaptive homomorphic deconvolution schemes for speech recognition in noise," in *IEEE Internat. Conf. Acoustics, Speech and Signal Processing*, vol. i, Apr 1994, pp. I/441–I/444 vol.1.

[34] C. S. Yip, S. H. Leung, and K. K. Chu, "Optimal root cepstral analysis for speech recognition," in *IEEE Internat. Symp. Circuits and Systems*, vol. 2, 2002, pp. II–173–II–176 vol.2.

[35] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[36] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 2, pp. 254 – 272, Apr. 1981.

[37] C. Avendano and H. Hermansky, "On the effects of short-term spectrum smoothing in channel normalization," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 4, pp. 372–374, 1997.

[38] N. R. Shabtai, B. Rafaely, and Y. Zigel, "The effect of reverberation on the performance of cepstral mean subtraction in speaker verification," *Applied Acoustics*, vol. 72, no. 2-3, pp. 124 – 126, 2011.

[39] D. Gelbart and N. Morgan, "Double the trouble: handling noise and reverberation in far-field automatic speech recognition." in *Internat. Conf. Spoken Language Processing*, 2002, pp. 2185–2188.

[40] R. Petrick, K. Lohde, M. Wolff, and R. Hoffmann, "The harming part of room acoustics in automatic speech recognition." in *INTERSPEECH*. ISCA, 2007, pp. 1094–1097.

[41] T. Kobayashi and S. Imai, "Spectral analysis using generalized cepstrum," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 32, no. 5, pp. 1087 – 1089, Oct. 1984.

[42] L. Nivanen, A. L. Méhauté, and Q. Wang, "Generalized algebra within a nonextensive statistics," *Rep. Math. Phys.*, vol. 52, no. 3, pp. 437 – 444, 2003.

[43] E. P. Borges, "A possible deformed algebra and calculus inspired in nonextensive thermostatistics," *Physica A,*, vol. 340, pp. 95–101, Sep. 2004.

[44] H.-G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems." in *INTERSPEECH*. Citeseer, 2005, pp. 2697–2700.

[45] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.