NFORMATION SYSTEM APPLICATION, Vol. 01, No. 02
ISSN 2503-166X

Data Suara Ucapan Vokal Bahasa Indonesia

[1]**Tjong Wan Sen**
[1]President University, Jl. Ki Hajar Dewantara, Cikarang Baru, Bekasi
[1]Fakultas Komputer
e-mail: [1] wansen@president.ac.id

***Abstract*— Automatic Speech Recognition (ASR) is a very useful technology to human being. Eventhough this technology has been developed for many years it still could not be used widely for daily activities. One of the problem that limits its usage is the diffrerences between languages. Once an ASR is trained using a language, that ASR could not properly recognize speech in another language. Bahasa Indonesia is one of the most used language in the world. There are more than 250 millions peoples use Bahasa Indonesia everyday. But the availability of ASR in Bahasa Indonesia is still limited. There are several speech databases in Bahasa Indonesia which are developed by big companies but they could not be accessed publicly. To further help the development of ASR in Bahasa Indonesia a speech database in Bahasa Indonesia must be developed. In this paper, an attempt to collect sound of speech vowel in Bahasa Indonesia is reported. Vowel sound from several major etnics of Bahasa Indonesia user which have big differences in sound are recorded. The speakers are divided into two groups by gender (male or female) and age (adult). The frequency characteristic for each group are investigated and the similarities are measured using correlation to further improve the quality.**

***Intisari*— Pengenalan ucapan otomatis adalah teknologi kecerdasan buatan yang sangat bermanfaat bagi manusia tetapi masih tetap belum dapat diimplementasikan secara luas dalam kehidupan sehari-hari. Selain disebabkan oleh karena keragaman cara pengucapan setiap individu yang sangat bervariasi, banyaknya bahasa yang berbeda-beda juga berperan besar dalam mengurangi potensi teknologi ini untuk digunakan secara umum. Hal tersebut disebabkan oleh karena basis data suara ucapan untuk suatu bahasa tertentu tidak dapat digunakan untuk mengenali ucapan dalam bahasa yang lain dengan baik secara langsung.**

Bahasa Indonesia adalah bahasa resmi negara Indonesia yang berpenduduk nomor empat terbesar di dunia. Pada saat ini belum banyak dikembangkan basis data suara ucapan manusia yang menggunakan Bahasa Indonesia dan basis data yang sudah ada yang dikembangkan oleh perusahaan besar pada umumnya tidak dapat diakses oleh publik dengan mudah. Hal tersebut mengurangi kemungkinan Pengenalan Ucapan Otomatis dengan Bahasa Indonesia untuk dapat dikembangkan lebih lanjut.

Dalam artikel ini dilaporkan pengembangan data ucapan vokal Bahasa Indonesia. Pengumpulan data dilakukan dengan cara merekam suara vokal dari berbagai etnis mayoritas yang ada di Indonesia yang memiliki perbedaan suara signifikan. Pengucap terbagi menjadi dua kelompok jenis kelamin (pria-wanita) dan usia (dewasa). Karakteristik respon frekuensi setiap rekaman dalam satu kelompok dibandingkan satu dengan lainnya menggunakan korelasi untuk penyempurnaan.

***Kata Kunci*—data suara ucapan, karakteristik frekuensi suara ucapan, pengenalan ucapan otomatis, suara ucapan vokal Bahasa Indonesia.**

I. PENDAHULUAN

Komunikasi suara merupakan cara yang murah, cepat dan sangat natural bagi manusia. Kemampuan ini didapat oleh setiap individu sejak lahir, dilatih secara terus menerus dan digunakan setiap hari sehingga pada umumnya setiap orang dapat melakukannya secara otomatis tanpa usaha yang berarti. Dengan adanya jaringan telekomunikasi dan komputer yang maju, jangkauan komunikasi suara dapat diperjauh hingga menjangkau seluruh dunia dan bahkan lebih. Sampai dengan detik ini, komunikasi suara melalui ucapan masih menjadi cara utama yang digunakan manusia. Mendominasi jauh diatas cara lain seperti mengetik dan menulis.

Teknologi elektronika pada umumnya dan komputer khususnya di berbagai bidang sudah mencapai tingkat yang cukup maju. Otomatisasi dilakukan hampir di semua komponen yang terlibat. Bagian yang masih dikerjakan secara manual, hari demi hari, semakin berkurang. Semua ditujukan terutama untuk kemudahan pengguna (ramah kepada pengguna atau *user friendly*). Meskipun dari segi yang lain seperti kecepatan dan efisiensi juga turut berperan penting. Terlebih lagi dengan adanya kemajuan di bidang komputer mulai dari *server* yang berkekuatan besar sampai kepada *smartphone* yang ringkas dan memiliki tingkat mobilitas tinggi.

Dalam peralatan elektronika dan komputer tersebut, alat input yang digunakan seperti *mouse*, *keyboard*, *touchscreen*, *keypad*, *touchpad*, *stylus* dan lain-lain belum ada yang bisa menyamai kemudahan dan keramahan atau bahkan menggantikan peran komunikasi suara ucapan. Oleh karena itu manusia telah senantiasa berusaha untuk mengembangkan teknologi pengenalan ucapan manusia sebagai alat input yang lebih baik. Peran alat input yang lain masih dipertahankan karena tidak semua bisa ditangani oleh komunikasi suara tetapi diharapkan porsi yang semakin besar bisa dicapainya oleh teknologi ini.

Pada saat ini, teknologi pengenalan ucapan otomatis yang dapat secara otomatis mengubah suara menjadi teks, telah dikembangkan selama beberapa dekade dan kemampuannya untuk memudahkan aktivitas manusia dalam kehidupan sehari-hari sudah terbukti. Namun teknologi ini belum dapat dimanfaatkan oleh secara luas oleh semua manusia. Hal tersebut disebabkan oleh beberapa hal seperti akurasi, kecepatan mesin dan basis data suara ucapan. Basis data suara

ucapan adalah salah satu komponen penting dalam teknologi ASR. Dari segi basis data, basis data suara ucapan dalam bahasa tertentu sudah pasti tidak dapat digunakan untuk mengenali suara ucapan dalam bahasa yang lain dengan baik. Meskipun untuk bahasa yang memiliki kemiripan tinggi masih dimungkinkan.

Salah satu basis data suara ucapan yang belum dikembangkan secara lengkap dan terpublikasi secara luas adalah basis data suara ucapan dalam Bahasa Indonesia. Beberapa sistem ASR sudah dapat mengenali ucapan dalam Bahasa Indonesia, tetapi basis data suara ucapannya tidak dapat diakses secara langsung tanpa perangkat lunak yang sudah ditentukan oleh penyedia layanannya. Di sisi lain, Bahasa Indonesia digunakan oleh sangat banyak pengguna. Hal tersebut dikarenakan penduduk Indonesia adalah keempat terbanyak di dunia. Juga karena Bahasa Indonesia serumpun atau termasuk dekat dengan Bahasa Melayu lainnya. Sehingga menambah potensinya dari sisi jumlah calon pengguna.

Pada dasarnya basis data suara ucapan adalah kumpulan dari rekaman suara ucapan dari banyak orang yang berbeda-beda yang mengucapkan banyak ucapan yang berbeda-beda pula. Perbedaan jenis kelamin, usia, gaya bicara dan aksen memiliki pengaruh yang besar disamping sumber perbedaan lainnya. Perekaman dan penyuntingan yang cukup dapat menyediakan informasi bagi basis data ucapan dalam suatu bahasa tertentu yang baik.

II. TEORI

Pada bagian ini akan dipaparkan teori yang mendasari langkah-langkah yang digunakan dalam pengembangan data suara vokal ucapan yang dilakukan.

A. Pengenal Ucapan Otomatis

Pengenalan ucapan otomatis mengambil suara ucapan sebagai masukan (input) melalui mikrofon dan menghasilkan tampilan teks di layar monitor, respon suara dari speaker, sinyal untuk mengendalikan perangkat elektronik dan lainnya sebagai keluaran (output). Pengenalan ucapan otomatis memiliki tiga komponen utama yaitu pengekstrak fitur, pengenal dan basis data ucapan [1]. Basis data ucapan memiliki tiga sub komponen yaitu model akustik, model kamus dan model bahasa. Pengekstrak fitur bertugas untuk mengubah suara yang diterima menjadi himpunan kode unik yang representatif. Pengenal bertugas untuk mencari kemiripan yang paling tinggi dari kode yang dihasilkan oleh pengekstrak fitur dengan kode yang tersimpan didalam basis data ucapan. Basis data ucapan bertugas untuk menyimpan semua kode unik yang dihasilkan pada saat sistem dalam proses pelatihan. Model akustik menyimpan data tentang satuan terkecil pengenalan seperti fonem, suku kata atau lainnya. Model kamus menyimpan informasi mengenai kata yang dapat dikenali yang merupakan kombinasi dari satuan terkecil pengenalan yang ada dalam model akustik. Model bahasa menyimpan informasi mengenai aturan tata bahasa yang berlaku yang mengatur susunan kompbinasi kata yang terdapat dalam model kamus.

B. Basis Data Suara Ucapan

Basis data suara ucapan yang terdiri atas tiga sub komponen yaitu model akustik, model kamus dan model bahasa, dihasilkan pada saat proses pelatihan [2]. Pada proses ini, untuk model akustik, pengenal ucapan pada awalnya menampung semua kode unik yang dihasilkan oleh pengekstrak fitur untuk setiap data hasil perekaman dan penyuntingan atau disebut juga data latih. Satu data rekaman menghasilkan satu fitur yang unik. Pada tahap berikutnya semua kode unik yang dihasilkan oleh data latih dengan kelas yang sama (misalnya fonem A yang diucapkan beberapa kali oleh seorang sumber yang sama atau beberapa suara fonem A yang dihasilkan oleh beberapa sumber yang berbeda) diproses lebih jauh untuk mendapatkan sebuah (atau beberapa) fitur yang dapat mewakili semua data latih dalam kelompok tersebut secara keseluruhan. Tergantung kepada bahasa yang digunakan atau batasan lainnya, model akustik basis data suara ucapan hanya memiliki kelas yang terbatas yaitu sejumlah satuan terkecil pengenalan yang ingin dikenali.

Pada model kamus, satuan terkecil pengenalan dalam model akustik dikombinasikan membentuk kata yang memiliki arti. Tidak semua kombinasi memiliki arti dan oleh karena itu tidak semua disimpan dalam model kamus. Pada umumnya model kamus terbatas pada jumlah kata yang dimiliki oleh suatu bahasa atau lebih kecil. Misalnya hanya kata yang terlibat pada dunia perbankan, kedokteran, percakapan ringan atau lainnya.

Pada model bahasa, semua kata yang ada dalam model kamus dikombinasikan untuk membentuk kalimat. Tidak semua kombinasi kata dapat membentuk kalimat yang baik yang memenuhi aturan tata bahasa bahasa yang bersangkutan. Kombinasi kata yang tidak memenuhi aturan tata bahasa yang bersangkutan tidak disimpan dalam model bahasa dan tidak dapat dikenali.

C. Vokal Bahasa Indonesia.

Masing-masing bahasa memiliki jumlah fonem yang tertentu. Fonem adalah unsur ucapan terkecil yang berbeda satu dengan lainnya. Kombinasi dari fonem membentuk kata atau kata yang berbeda tersusun dari fonem yang berbeda pula. Fonem pada umumnya terbagi ke dalam dua kelompok besar yaitu vokal dan konsonan. Vokal memiliki suara yang lebih jelas dn tegas sedangkan konsonan kebalikannya. Perbedaan ini disebabkan oleh karena berbedanya unsur frekuensi dan besar energi yang terkandung didalam fonem tersebut.

Jumlah fonem dalam Bahasa Indonesia diatur oleh Tata Bahasa Indonesia [3]. Ada lima buah fonem vokal utama yaitu A, E, I, O, dan U. Seluruh kata yang terdapat dalam kamus Bahasa Indonesia disusun oleh kombinasi fonem vokal sebagai unsur utama. Perbedaan lebih lanjut diberikan oleh fonem konsonan yang berada diawal, akhir dan antara fonem vokal tersebut. Karena karakteristik tersebut fonem vokal

menjadi lebih mudah dikenali dibandingkan dengan fonem konsonan.

D. Frekuensi Suara Ucapan

Dari sudut pandang frekuensi, suara ucapan dapat direpresentasikan sebagai represntasi dari sebuah himpunan nilai yang bersesuaian dengan masing-masing komponen frekuensi. Sampel dari rekaman suara dalam domain waktu dapat ditransformasikan ke dalam domain frekuensi melalui Transformasi Fourier [4]. Pada domain frekuensi sinyal rekaman suara *F(u)*, yang merupakan hasil dari Transformasi Fourier sinyal *f(x)* dari domain waktu, direpresentasikan sebagai jumlah tidak terhingga dari penjumlahan gelombang sinus dan cosinus dari semua frekuensi seperti ditunjukkan oleh (1) dan (2).

$$F(u) = \int_{-\infty}^{\infty} f(x)\, e^{-i2\pi ux} dx \quad (1)$$

atau

$$F(u) = \int_{-\infty}^{\infty} f(x)\, (cos\ 2\pi ux - i\ sin\ 2\pi ux)\, dx \quad (2)$$

Fungsi *F(u)* adalah fungsi kontinu dalam domain frekuensi yang merupakan bilangan kompleks. Fungsi *f(x)* adalah fungsi kontinu dalam domain waktu dengan variabel riil *x* yang merupakan representasi dari sinyal rekaman suara dalam domain waktu. Variabel *u* adalah variabel frekuensi dalam domain frekuensi.

Untuk bekerja dengan sinyal diskrit maka diperlukan Transformasi Fourier Diskrit. Dengan mengambil sampel dari fungsi *f(x)* secara periodik maka hasil Transformasi Fourier yang bersesuaian dapat diperoleh melalui (3).

$$F(u) = \frac{1}{N} \sum_{x=0}^{N-1} f(x) e^{-i2\pi ux/N} \quad (3)$$

Variabel *N* adalah banyaknya sampel diskrit yang diambil dari fungsi *f(x)*. Pada penelitian ini digunakan salah satu bentuk khusus dari Transformasi Fourier Diskrit yaitu Cooley-Tukey Fast Fourier Transform (FFT) yang dapat mereduksi kebutuhan komputasi dari N^2 menjadi $N \log_2 N$ [5].

E. Koefisien Korelasi

Koefisien Korelasi dari dua variabel acak menunjukkan tingkat ketergantungan linier diantara keduanya. Jika ada observasi sebanyak *N* maka Koefisien Korelasi ρ dapat ditentukan oleh (4) [6]. Barisan *A* dan *B* adalah barisan pasangan observasi yang bersesuaian, μ_A dan σ_A adalah standar deviasi dari barisan *A* dan μ_B dan σ_B adalah standar deviasi dari barisan *B*.

$$\rho(A,B) = \frac{1}{N-1} \sum_{k=1}^{N} \left(\frac{A_i - \mu_A}{\sigma_A}\right)\left(\frac{B_i - \mu_B}{\sigma_B}\right) \quad (4)$$

Jika pasangan nilai dalam barisan *A* dan *B* yang bersesuaian memiliki ketergantungan linier yang tinggi, misalnya jika nilai dalam barisan *A* bertambah maka nilai dalam barisan *B* pun bertambah, maka nilai dari Koefisien Korelasi akan mendekati 1. Sebaliknya jika nilai dalam barisan *B* tidak tergantung pada nilai yang bersesuaian dalam barisan *A* makan nilai Koefisien Korelasi akan mendekati 0.

III. METODA PENELITIAN

Tahap pertama dalam penelitian adalah tahap perekaman suara vokal dari berbagai sumber pengucap. Masing-masing sumber ucapan akan mengucapkan lima puluh vokal acak secara terpisah. Untuk keperluan itu digunakan perangkat lunak yang berfungsi untuk a) memulai rekaman, b) menampilkan vokal yang harus diucapkan pengucap, c) menghentikan rekaman setelah waktu tertentu, d) menyimpan hasil rekaman ke dalam format *file wav* dengan nama yang sudah ditentukan dan e) mengulang proses dari awal (a). Untuk memberikan waktu yang cukup bagi pengucap, lama untuk satu rekaman diberikan waktu 2 detik. Suara direkam secara mono dengan frekuensi sampling 44.100 Hz. Bit kuantisasi yang digunakan adalah 16 bit. Perekaman dilakukan di ruangan kerja biasa bukan di studio atau laboratorium khusus untuk merekam.

Tahap selanjutnya adalah tahap verifikasi dan penyuntingan. Pada tahap verifikasi hasil rekaman dipisahkan berdasarkan vokal yang diucapkan (berdasarkan nama *file*). Kemudian *file-file* tersebut diputar ulang dan didengarkan secara manual untuk memastikan tidak ada suara vokal yang tidak sesuai dengan labelnya atau rusak. Pada tahap penyuntingan potongan hasil rekaman diekstrak dari masing-masing file selama 0,128 detik atau setara dengan 5.632 sampel berdasarkan kumpulan sampel (sebanyak 2.048 sampel) dengan energi yang paling besar. Jumlah 2.048 sampel atau setara dengan 0,046 detik tersebut disesuaikan dengan masukan untuk FFT. Sampel lainnya merupakan hasil pergeseran 256 sampel sebanyak 7 bagian sebelum kumpulan sampel dengan energi terbesar dan 7 bagian ke kanan setelahnya.

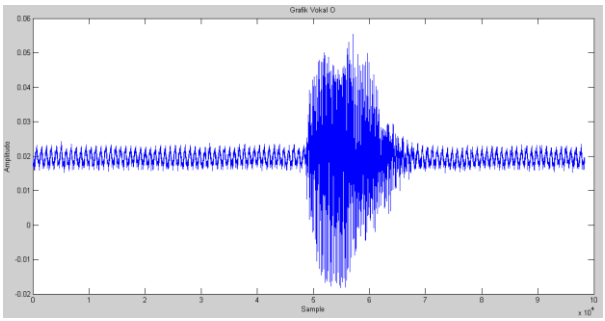
Tahap yang ketiga adalah tahap dimana seluruh hasil dari tahap verifikasi dan penyuntingan ditransformasi ke domain frekuensi dengan bantuan FFT. Dengan masukan FFT sebesar 2.048 sampel maka akan diperoleh 1.024 kelompok frekuensi (yang direpresentasikan oleh bilangan kompleks) yang bersesuaian dengan respon frekuensi 0 sampai dengan 22.050 Hz (setengah dari frekuensi sampling 44.100 Hz). Kumpulan respon frekuensi tersebut inilah yang disimpan sebagai data suara vokal Bahasa Indonesia. Kumpulan ini dapat diproses lebih lanjut menjadi fitur suara ucapan vokal Bahasa

Indonesia dengan berbagai cara baik yang sudah umum maupun yang baru dikembangkan.

Setelah kumpulan respon frekuensi suara vokal Bahasa Indonesia dihasilkan, penelitian dilanjutkan dengan melakukan analisis kemiripan antara satu dengan lainnya dengan menggunakan koefisien korelasi. Pada tahap ini sebuah respon frekuensi dianggap sebagai suatu barisan observasi (barisan *A*) dan sebuah respon frekuensi lainnya sebagai barisan observasi yang bersesuaian (barisan *B*). Koefisien korelasi yang tidak logis, misalnya nilai yang rendah untuk dua respon frekuensi dari suara rekaman vokal yang sama, digunakan untuk melakukan pengecekan kembali ke sumbernya. Jika ada kesalahan dalam proses maka akan diperbaiki atau bahkan diulang. Akan tetapi jika memang proses sudah benar maka respon frekuensi tersebut akan dipisahkan sekaligus diberi informasi tambahan yang sesuai untuk keperluan di masa depan. Dengan cara seperti itu diharapkan kumpulan data suara ucapan vokal Bahasa Indonesia yang dihasilkan menjadi lebih baik dan dapat dengan mudah dimanfaatkan lebih lanjut untuk sistem pengenalan ucapan otomatis dalam Bahasa Indonesia. Perbaikan dan penyempurnaan ini dilakukan secara iteratif untuk mendapatkan hasil yang sebaik mungkin.

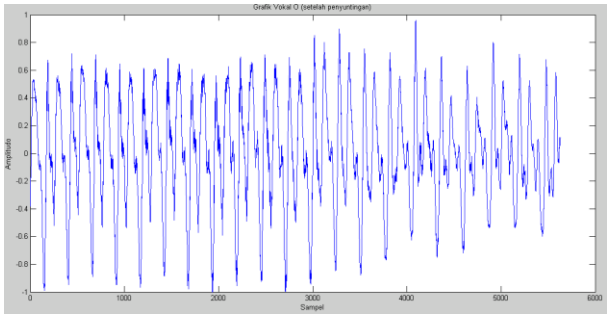
IV. HASIL DAN PEMBAHASAN

Hasil dari tahap perekaman adalah berupa file dalam format *wav* yang berisikan masing-masing satu vokal dari seorang pengucap. Jika ditampilkan dalam bentuk grafik dengan bantuan perangkat lunak MATLAB adalah seperti ditunjukkan oleh Gbr 1.



Gbr. 1 Bentuk gelombang salah satu vokal O yang direkam.

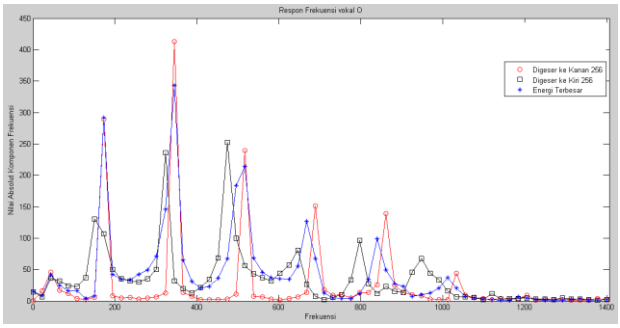
Seperti terlihat dalam Gbr. 1, ada sekitar 98.000 sampel (dalam rekaman berdurasi 2 detik dengan frekuensi sampling 44.100 Hz) dengan besar amplitudo masing-masing. Di bagian awal dan akhir rekaman tidak terdapat informasi mengenai vokal yang bersangkutan dan bisa dihilangkan (pada tahap penyuntingan). Pengolahan hasil rekaman lebih lanjut ditujukan untuk maksud tersebut selain juga untuk menyamakan hasil rekaman secara keseluruhan. Dua sumber perbedaan yang utama adalah keras-pelannya suara yang dihasilkan pengucap dan karakteristik mikrofon dan kartu suara yang digunakan (seperti terlihat dalam Gbr. 1, titik tengah amplitudo yang tidak berada di titik nol merupakan pengaruh dari kartu suara yang digunakan untuk merekam).



Gbr. 2 Bentuk gelombang vokal O dalam Gbr. 1 setelah melalui penyuntingan.

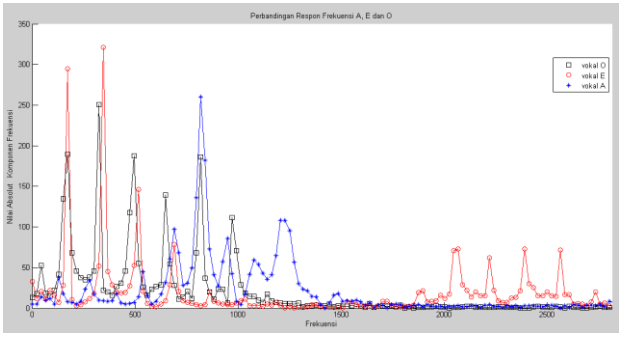
Hasil dari tahap penyuntingan rekaman potongan suara ucapan yang memiliki kandungan energi paling besar. Salah satu contoh hasil penyuntingan dari sebuah suara vokal hasil rekaman (yaitu vokal O yang ditunjukkan oleh Gbr. 1) ditunjukkan oleh Gbr. 2.

Seperti ditunjukkan oleh Gbr. 2, terdapat sekitar 5.600 sampel beserta amplitudonya masing-masing. Jumlah tersebut disesuaikan dengan jumlah sampel input untuk satu kali Transformasi Fourier (yaitu 2.048) dengan pergeseran 256 sampel antara dua transformasi yang berurutan. Sehingga dari setiap hasil penyuntingan terdapat 15 kelompok input bagi Transformasi Fourier.



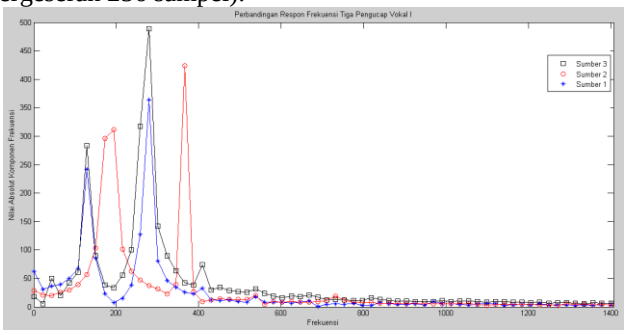
Gbr. 3 Respon frekuensi dari potongan vokal O dalam Gbr. 2.

Hasil dari tahap Transformasi Fourier (dalam hal ini menggunakan FFT) adalah respon frekuensi dari himpunan sampel yang bersesuaian. Gbr. 3 menunjukkan tiga respon frekuensi dari tiga himpunan sampel dari vokal O yang ditunjukkan oleh Gbr. 2. Himpunan sampel yang memiliki energi paling tinggi (bintang), himpunan sampel yang digeser 256 sampel ke kiri (lingkaran) dan himpunan sampel yang digeser ke kanan 256 sampel (kotak).



Gbr. 4 Perbandingan respon frekuensi vokal A E dan O dari seorang pria dewasa.

Seperti ditunjukkan oleh Gbr. 3, pergeseran himpunan sampel memiliki pengaruh kepada respon frekuensi. Selain perubahan frekuensi (sumbu x) juga besar absolut yang bersesuaian dengan komponen frekuensi tersebut (sumbu y). Hal ini menunjukkan untuk vokal yang sama yang diucapkan oleh pengucap yang sama dapat memiliki respon frekuensi yang berbeda jika diambil dari waktu yang berbeda (pergeseran 256 sampel).



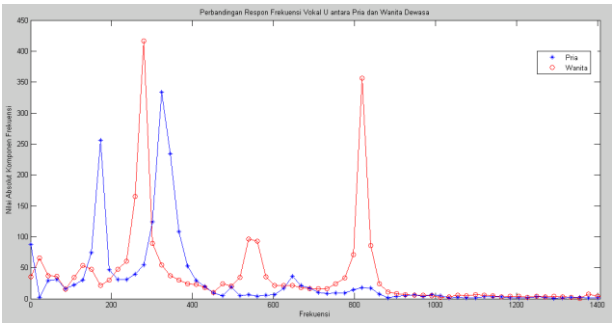
Gbr. 5 Respon frekuensi Vokal I dari tiga sumber yang berbeda.

Setelah semua respon frekuensi diperoleh, tahap selanjutnya adalah mengukur kemiripan antar respon frekuensi tersebut. Perbandingan dilakukan terhadap respon frekuensi dari orang yang sama dengan vokal yang berbeda dan terhadap vokal yang sama yang diucapkan oleh orang yang berbeda. Sebagai contoh perbandingan dari respon frekuensi seorang pengucap pria dewasa yang mengucapkan vokal A, E dan O ditunjukkan oleh Gbr. 4.

Terlihat pada Gbr. 4 respon frekuensi vokal E (lingkaran) memiliki kemiripan dengan vokal O (kotak) pada frekuensi rendah dan perbedaan yang cukup signifikan terlihat pada frekuensi yang lebih tinggi. Sedangkan vokal A menunjukkan perbedaan yang signifikan terhadap kedua vokal lainnya. Hal yang sama terjadi juga pada sumber-sumber ucapan yang lain dengan tingkat perbedaan yang bervariasi. Contoh berikutnya ada perbandingan respon frekuensi dari tiga sumber ucapan vokal I yang berbeda. Perbedaan respon frekuensi ditunjukkan oleh Gbr. 5.

Seperti ditunjukkan oleh Gbr. 5, sumber vokal I yang kedua (lingkaran) memiliki perbedaan yang cukup signifikan dari sumber 1 (bintang) dan 3 (kotak). Akan tetapi jika dibandingkan dengan respon frekuensi vokal A, E dan O pada Gbr. 3 tetap terlihat perbedaan yang signifikan. Pada bagian selanjutnya dilakukan perbandingan dari suara vokal yang sama dari sumber ucapan pria dewasa dan wanita dewasa. Perbandingan tersebut ditunjukkan oleh Gbr. 6.

Terlihat perbedaan yang cukup signifikan antara respon frekuensi vokal U yang diucapkan oleh pengucap pria dengan wanita dewasa. Komponen frekuensi wanita (lingkaran) terlihat lebih tinggi dari pria (bintang).



Gbr. 6 Perbandingan respon frekuensi vokal U pria dan wanita dewasa.

Untuk memberikan perbandingan kemiripan yang lebih lengkap antara semua respon frekuensi untuk selanjutnya digunakan koefisien korelasi. Sebagian dari koefisien korelasi tersebut ditunjukkan oleh Tabel I sampai dengan Tabel III.

TABEL I
KOEFISIEN KORELASI PERGESERAN SAMPEL

	X-7	X-4	X-2	X	X+2	X+4	X+7
X-7	1	0,98	0,97	0,91	0,90	0,85	0,86
X-4	0,98	1	0,97	0,90	0,88	0,82	0,82
X-2	0,97	0,97	1	0,97	0,94	0,87	0,83
X	0,91	0,90	0,97	1	0,96	0,90	0,83
X+2	0,90	0,88	0,94	0,96	1	0,97	0,91
X+4	0,85	0,82	0,87	0,90	0,97	1	0,93
X+7	0,86	0,82	0,83	0,83	0,91	0,93	1

Seperti ditunjukkan oleh Tabel I, pergeseran sejauh 256 sampel ke kanan dan kiri dari sampel suara yang memiliki kandungan energi paling besar mempengaruhi kemiripan komponen frekuensi. Koefisien korelasi memiliki nilai yang berkisar dari 0,82 (terkecil) sampai dengan 0,98 (terbesar). Pada umumnya semakin besar pergeseran semakin berbeda respon frekuensi yang diperoleh.

TABEL III
KOEFISIEN KORELASI SATU VOKAL DARI LIMA SUMBER

	A1	A2	A3	A4	A5	A6*
A1	1	0,58	0,34	0,53	0,49	0,46
A2	0,58	1	0,25	0,53	0,40	0,28
A3	0,34	0,25	1	0,41	0,40	0,38
A4	0,53	0,53	0,41	1	0,40	0,19

A5	0,49	0,40	0,40	0,40	1	0,28
A6*	0,46	0,28	0,38	0,19	0,28	1

*A6 berasal dari kelompok yang berbeda yaitu wanita dewasa.

Dari Tabel II dapat dilihat bahwa koefisien korelasi antara dua sumber ucapan yang berbeda walaupun vokal yang diucapkan sama (vokal A) dan berasal dari kelompok yang sama (pria dewasa) memiliki nilai yang tidak terlalu tinggi. Nilai koefisien korelasi berkisar antara 0,58 (antara A1 dengan A2) sebagai yang tertinggi dan 0,25 (antara A2 dengan A3). Tetapi nilai tersebut masih lebih tinggi dari 0,19 (antara A4 dengan A6*) dimana A6* adalah suara vokal A yang diucapkan oleh seorang wanita dewasa.

TABEL IIIII
KOEFISIEN KORELASI LIMA VOKAL DARI SATU SUMBER

	A	**E**	**I**	**O**	**U**
A	1	0,11	0,08	0,44	0,08
E	0,11	1	0,72	0,42	0,41
I	0,08	0,72	1	0,54	0,52
O	0,44	0,42	0,54	1	0,46
U	0,08	0,41	0,52	0,46	1

Perbedaan respon frekuensi dari vokal yang berbeda yang diucapkan oleh orang yang sama ditunjukkan oleh Tabel III. Pada umumnya respon frekuensi antara dua vokal yang berbeda menunjukkan perbedaan yang signifikan dengan nilai berkisar pada 0,08 (terkecil) dan 0,54 (terbesar). Khususnya pada vokal E dan I terdapat kemiripan yang signifikan yang ditunjukkan oleh koefisien korelasi 0,72. Hal tersebut menunjukkan bahwa dengan menggunakan pengukuran korelasi vokal E dan I memiliki kemiripan yang tinggi untuk sebagian sumber ucapan meskipun secara bunyi suara cukup berbeda.

V. PENUTUP

Kumpulan data suara vokal Bahasa Indonesia yang berisikan respon frekuensi hasil Transformasi Fourier dari suara ucapan lima vokal utama dalam Bahasa Indonesia telah dihasilkan. Sumber ucapan diambil dari banyak pengucap yang mewakili etnis mayoritas yang ada di Indonesia dan memiliki bunyi suara ucapan yang berbeda secara signifikan. Sumber ucapan terbagi menjadi dua kelompok yaitu pria dewasa dan wanita dewasa. Kumpulan respon frekuensi ini dapat dengan mudah dimanfaatkan sebagai masukan untuk modul pengekstraksi fitur dalam sistem pengenalan ucapan otomatis menggunakan Bahasa Indonesia.

Sebagai pengembangan lebih lanjut penelitian ini dapat ditingkatkan dengan menambahkan kelompok dari sisi usia yaitu kelompok anak laki-laki dan perempuan atau bahkan lansia. Pengukuran kinerja juga dapat dilanjutkan dengan memanfaatkan metoda *clustering* untuk pengelompokkannya secara *unsupervised* atau *Support Vector Machine* (*supervised*). Juga dapat dilakukan investigasi kemungkinannya untuk dilakukan reduksi dimensi menggunakan *Principle Component Analysis* (PCA).

UCAPAN TERIMA KASIH

Ucapan terima kasih yang sebesar-besarnya diberikan kepada Lembaga Riset dan Pengabdian Masyarakat xxxxxxxxx xxxxxxxxx yang telah memberikan bantuan yang sangat berarti bagi penelitian ini. Juga kepada semua pihak yang terlibat secara langsung maupun tidak langsung yang tidak dapat disebutkan satu persatu disini disampaikan penghargaanyang setinggi-tingginya.

REFERENSI

[1] S. Furui, "Automatic speech recognition and its application to information extraction", Association for Computational Linguistics, hal. 11–20, 1999.

[2] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, vol. 77, no. 2. hal. 257–286, 1989.

[3] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A. M. Moeliono, Tata Bahasa Baku Bahasa Indonesia, Edisi Ketiga, Jakarta: Balai Pustaka, 2014.

[4] Y. Yoo, Tutorial on Fourier theory, (2001).

[5] P. Duhamel and M. Vetterli, "Fast Fourier transforms: a tutorial review and a state of the art", Signal Processing, vol. 19, hal. 259–299, 1990.

[6] (2016) The MathWorks Documentation website. [Online], http://www.mathworks.com/help/matlab/ref/corrcoef.html, tanggal akses: 02 Juni 2016.