Peningkatan Akurasi Kualitas Suara Hasil Transformasi Dengan Pemetaan Spektral Pada Kondisi Supervisi dan Nonsupervisi

Driszal Fryantoni
Pusat Penelitian Informatika- LIPI
driszal@informatika.lipi.go.id

Abstract

This paper describes quality of speech transformation using spectral mapping analysis synthesis and power spectral envelope in supervised and nonsupervised conditions. Speaker spectral input can be mapped by interpolation and fuzzy code to get a speech reference which will be used to identify characteristics of speaker speech. From the speaker identification evaluation results, based on comparison between transformed speech and speech reference, the trial result proved that the percentage of man to man speech transformation for two methods above is 84%, and man to women speech transformation identification ratio is 70% in supervised condition.

Keywords: Supevised and nonsupervised spectral maping, power spectral enveloped, fuzzy code.

Abstrak

Dalam tulisan ini dijelaskan mengenai kualitas transformasi suara dengan analisa sintesa pemetaan spectral dan amplop power spectral, pada kondisi supervisi dan non-supervisi. Pemetaan spectral suara pembicara dilakukan untuk mendapatkan referensi suara, yang akan menjadi acuan dalam melakukan identifikasi karakteristik suara pembicara, dengan memakai interpolasi dan pengkodean fuzzy. Dari hasil evaluasi percobaan identifikasi suara pembicara, dapat diketahui bahwa rata-rata prosentase identifikasi, transformasi suara laki-laki ke suara laki-laki untuk kedua cara diatas adalah 84%, sedangkan transformasi suara laki-laki ke perempuan diperoleh prosentase rata-rata identifikasi sebesar 70% pada kondisi supervisi.

Kata kunci: Pemetaan spektral supervisi dan non supervisi, Amplop spektral power, pengkodean fuzzy.

1. Pendahuluan

Untuk mewujudkan *man machine interface* secara alami, pengontrolan karakteristik *suara pembicara* pada suara buatan, merupakan salah satu pekerjaan yang sangat penting. Akan tetapi, transformasi kualitas suara agar menjadi suara pembicara independen juga merupakan bidang penelitian yang sangat ditunggu-tunggu perkembangannya. Seperti output suara buatan dari telepon penterjemah otomatis dan bilingual suara dalam siaran televisi, yang dapat menghasilkan kualitas suara seperti suara asalnya pembicara.

Tinggi rendahnya kualitas suara pembicara, banyak disebabkan oleh perbedaan secara anatomis (*vocal cord*, dan *vocal tract* sebagai perangkat penghasil suara) dan secara akustikal (*formant frequency, kemiringan spectral serta fundamental frequency rata-rata)*, yang menjadi parameter-parameter penting untuk persepsi karakteristik suara pembicara [1].

Selama ini, salah satu metode yang banyak dilakukan untuk meningkatkan kualitas suara hasil transformasi, adalah dengan cara mengubah-ubah parameter akustiknya [2], akan tetapi kualitas hasil transformasi suara tidak memuaskan, hal ini disebabkan oleh permasalahan sebagai berikut. Yang pertama adalah tingkat akurasi pemetaan amplop spektral yang masih kurang, terutama pemetaan dengan menggunakan kuantisasi vektor yang dipengaruhi oleh rata-rata kesalahan kuantisasi *(error rate quantization)* [3]. Selanjutnya, dalam sistem analysis sinthesis LPC (*Linear Prediction Code*), tidak dapat menampilkan karakteristik secara detail pada amplop spektral sumber suara *(spectral envelope speech source)* [4][5].

Untuk meningkatkan tingkat akurasi transformasi kualitas suara dan perbaikan mutunya yang merupakan tujuan dari penelitian ini, pertama-tama kami akan perkenalkan metode supervisi dalam melakukan pemetaan, dengan meminimalisasi error antara spektral transformasi suara dan spektral referensi suara pembicara. Kemudian dilakukan pengamatan kualitas transformasi suara dalam sistem *analysis sinthesis power spectral envelope*.

Selain itu dilakukan juga pengamatan hasil analisa memakai pemetaan fuzzy dan *speaker adaptation non-supervisi,* dengan minimalisasi fungsi target fuzzy dari spektral codebooks. Metode konversi non-supervisi ini merupakan metode yang sangat mudah diaplikasikan, karena tidak memerlukan pengucapan yang sama dengan referensi suara dan input suara pembicara.

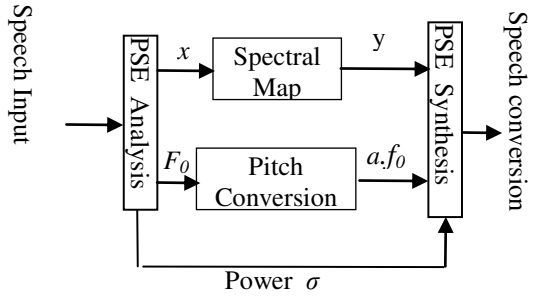
2. Sistem analisis sintesis transformasi kualitas suara

Struktur dari sistim analisis sintesis transformasi kualitas suara ditunjukkan pada gambar 1. Dari gambar 1 pada bagian analisis, dilakukan ekstraksi terhadap amplop spektral input suara (PSE/Power Spectral

Envelope), pitch (F0), dan power (σ). Dengan tidak merubah power disetiap frame, dilakukan pemetaan amplop spektral referensi suara. Agar nilai rata-rata pitch frekwensi sama dengan referensi suara, dilakukan perubahan frekwensi pitch secara proposional. Selanjutnya pada bagian sintesis, dengan menggunakan power dan parameter yang telah dikonversi dilakukan re-sintesis hasil konversi suara.

2.1 Bagian analysis

Agar struktur *pole-zero* amplop spektral karakteristik *vocal tract* dan *glottal source* dapat diekstrak dengan baik, kami menggunakan PSE analisis yang telah diperbaiki untuk estimasi amplop spektral [6].



Gambar 1. Struktur Sistem Analisis-Synthesis

Pertama, ekstraksi fundamental frekwensi F_0 terhadap gelombang suara yang telah disampling dengan sampling frekwensi F_s. Untuk menentukan bersuara/tidak bersuaranya pengucapan mamakai metode *cepstrum* menggunakan *N point* FFT. Lalu, dilakukan ekstraksi logaritma barisan sampling power spectral envelope $log|S(k_n)|^2$ *(n = 1, 2, ..., K)* dari logaritma power spektral $log|S(k)|^2$ *(k = 0, 1, ..., N-1)*. Sampling frekwensi barisan sampling PSE f_n, merupakan perkalian integral fundamental frekwensi F_0, dimana nilai maksimum terbesar pada sampling frekwensi barisan sampling PSE $f_n=k_nF_s/N$ memenuhi persamaan berikut :

$$nF_0 - F_0/2 < f_n < nF_0 + F_0/2 \qquad (1)$$

Nilai maksimum terbesar adalah jarak sampling yang lebih besar dari $1.5F_0$ dan ditetapkan sebagai sampling point. Estimasi barisan logaritma amplop spectral power$\{C_n\}$, dapat dituliskan dalam cosine series orde *ke-q* dengan metode least square, seperti persamaan di bawah ini,

$$log|H(k)|^2 = 2\sum C_n cos(2\pi nk/N) \qquad (2)$$

Dengan *normal equation*, $\{C_n\}$ dapat ditampilkan menjadi seperti persamaan di bawah ini,

$$\sum A_{mn}C_n=\sum cos(2\pi mk_n/N)ln|S(k_n)| \qquad (3)$$
$$(m = 0, 1, ..., q)$$

Dimana,

$$A_{mn} = \sum cos(2\pi mk_n/N)cos(2\pi nk_n/N) \qquad (4)$$

Untuk bagian bersuara, pada saat $K<q$, maka pada persamaan (3) menjadi $q=K$ di orde ke $K+1$ dan koefisien cepstrum PSE lainnya ditetapkan sama dengan nol. Sedangkan pada bagian tidak bersuara, F_0 dalam persamaan (1) ditetapkan *60 Hz* dan dilakukan estimasi cepstrum PSE seperti di atas. Kemudian dari hasil pengamatan analisa data, dilakukan perbaikan pada frame yang mengandung pitch error lalu dilakukan analisis ulang kembali.

2.2 Bagian synthesis

Pada bagian ini, dilakukan inverse mel-cepstrum PSE untuk mendapatkan PSE cepstrum frekwensi secara linear, agar dapat ditentukan amplitudo amplop spektral $|H(k)|$ memakai FFT dan eksponensial.

Untuk bagian bersuara, impulse respon gelombang dari hasil inverse FFT digunakan untuk menetapkan zero-phase dalam penyusunan deret impulse fundamental priodik satuan. Setelah power disamakan dengan gain input maka hasil output dari scalling amplitudo ditetapkan sebagai output suara. Untuk bagian tidak bersuara, adalah dengan cara mengatur impulse response yang telah ditentukan dengan memberikan gausian noise.

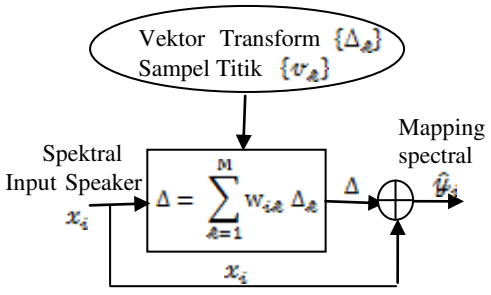
3. Metode pemetaan supervisi

Dalam bab 3 ini, akan kami uraikan secara singkat mengenai beberapa cara yang dipakai dalam metode pemetaan supervisi.

3.1 Interpolasi perbedaan spektral

Salah satu cara yang dipakai untuk memetakan input spektral pembicara x_i, menjadi spektral target pembicara $\hat{y}_i$ adalah dengan menggunakan persamaan interpolasi pada perbedaan spektral.

Pertama-tama, tentukan sembarang titik sampel v_k sebanyak M buah, dalam domain input spektral pembicara. Dengan menggunakan persamaan (5) dan (6), dapat dilakukan transformasi estimasi selisih antara vektor transformasi Δ_k (k = 1, …, M) dan target pembicara pada titik sampel v_k, sebagai ilustrasinya seperti ditunjukan pada Gambar 2.



Gambar 2. Pemetaan spektral supervisi

$$\hat{y}_i = x_i + \sum_{k=1}^{M} W_{ik}\Delta_k \qquad (5)$$

$$W_{ik} = \frac{\|x_i - v_k\|^{-p}}{\sum_{k=1}^{M} \|x_i - v_k\|^{-p}} \qquad (6)$$

p adalah parameter interpolasi yang mengontrol kontinyuitas pemetaan spektral. Selisih vektor pada y_i untuk $p = 0$ ditetapkan sebagai rata-rata vektor Δ_k Untuk $p \to \infty$, selisih rata-rata vektor tiap bagian pada jarak minimum $\{v_k\}$ adalah Δ_k.

3.2 Estimasi selisih vektor

Hubungan antara frame $\{P_r\} = \{i(r), j(r)\}$ $(r = 1 \sim L)$ dan selisih vector $\{\Delta_k\}$ diantara deret spectral $\{y_i\}$ target speaker, dan deret spectral $\{\hat{y}_i\}$ input speaker setelah pemetaan, dapat dituliskan sebagai fungsi target seperti berikut:

$$\xi(\{P_r\}, \{\Delta_k\}) = \sum_{r=1}^{L} \left\| \hat{y}_{i(r)} - y_{j(r)} \right\|^2 \qquad (7)$$

Dengan membuat minimum nilai pada persamaan (7), maka $\{P_r\}$ dan $\{\Delta_k\}$ dapat diestimasi, dimana L adalah jumlah total kisi-kisi point pada DP (Distance Point) Path, hingga dapat dituliskan kembali dalam persamaan normal seperti berikut ini,

$$\sum_{k=1}^{M} W_{rk} \Delta_k = \sum_{i=1}^{L} w_{i(i)r} \{ y_{j(i)} - x_{i(i)} \} \qquad (8)$$
$$(r = 1, ..., M)$$

Dan

$$W_{rk} = \sum_{i=1}^{L} w_{i(i)r} w_{i(i)k} \qquad (9)$$

Kata-kata komplek dapat ditraining dengan cara menghitung DP per-kata dari masing-masing kata dan kata yang berkelanjutan ditetapkan sebagai satu frame $\{P_r\}$.

3.3 Pemetaan dengan *codebooks*

Dengan mengasumsikan deretan sampel titik $\{v_k\}$ dalam spektral input pembicara sebagai Codebooks dan pada kondisi tidak ada interpolasi selisih spectral $(p \to \infty)$, maka persamaan (6) dapat ditulis kembali seperti berikut, dimana k menjadi kode quantisasi x_i

$$\lim_{p \to \infty} w_{ik} = \delta_{ik} \qquad (10)$$

Dengan melambangkan kode quantisasi $x_{i(l)}$ dengan lambang $c(l)$, maka dari persamaan (8), (9), (10) kita dapat menuliskan kembali persamaan (5) menjadi,

$$\hat{y}_i = \frac{\sum_{i=1}^{L} \delta c(l)_k \, y_{j(l)}}{\sum_{i=1}^{L} \delta c(l)_k} \qquad (11)$$

Dengan menggunakan pengkodean, dapat dikatakan bahwa pemetaan spektral x_i merupakan rata-rata dari spektral target pembicara $y_{j(l)}$, seperti hubungan pada persamaan (11).

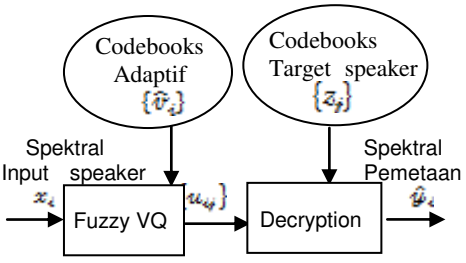
4. Metode pemetaan unsupervisi

Tahapan pemetaan spektral secara unsupervisi ditunjukkan pada Gambar 3, pertama-tama, melakukan kuantisasi vektor fuzy pada sembarang input spektral pembicara x_i dengan memakai acuan adaptif codebooks untuk mendapatkan class function. Lalu, dilakukan deskripsi terhadap class function tadi dengan memakai acuan codebooks referensi pembicara, untuk mendapatkan spektral transformasi pemetaan. Pendeskripsian dilakukan dengan memakai persamaan (12), seperti dibawah ini.

$$\hat{y}_j = \sum_{u_{ij} > u_r} \left[(u_{ij})^F \cdot z_i \right] \Big/ \sum_{u_{ij} > u_r} (u_{ij})^F \qquad (12)$$

Penjumlahan pada persamaan (12), mengandung arti jumlah dari *class function* yang memenuhi $u_{ij} > u_r$, dan diasumsikan $u_r = 1/(code\ size)$.

Dari penjelasan di atas, metode supervisi adalah pemetaan spektral yang input spektralnya mengalami tranformasi sendiri, sedangkan metode unsupervisi adalah pemetaan spektral yang terjadi dari spektral target speaker [7].



Gambar 3. Pemetaan spektral unsupervisi

5. Evaluasi akurasi pemetaan spektral

5.1 Evaluasi pemetaan supervisi

Evaluasi ini bertujuan untuk memeriksa pengaruh parameter interpolasi p dan jumlah titik sampel M, terhadap tingkat akurasi pemetaan spektral secara supervisi. Dalam evaluasi ini, kami menggunakan database suara yang terdiri dari kata-kata hasil pengucapan oleh 2 orang laki-laki (m_1, m_2) dan 2 orang perempuan (f_1, f_2) , dengan kondisi analisis suara seperti yang ditunjukkan pada Tabel 1.

Kata nomor 1 ~ 100 (jumlah frame 6000~7000) dalam database suara dijadikan sebagai training kata dan kata nomor 101 ~130 dijadikan sebagai evaluasi suara.

Lalu dilakukan estimasi vektor transformasi $\{\Delta_k\}$ pada 2 kelompok kombinasi transformasi suara pembicara (suara laki-laki →suara laki-laki dan suara laki-laki →suara perempuan).

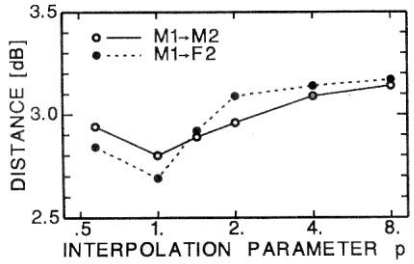
Tingkat akurasi pemetaan transformasi deret spektral $\{y_i\}$ ke deret spektral target $\{\hat{y}_i\}$, dapat kita evaluasi dengan memakai distance spectral, seperti persamaan di bawah ini,

$$D = 10 \cdot \log 10^e \cdot \sqrt{\frac{2}{L}\sum_{r=1}^{L}\left\|\hat{y}_{i(r)} - y_{j(r)}\right\|^2} \qquad (13)$$

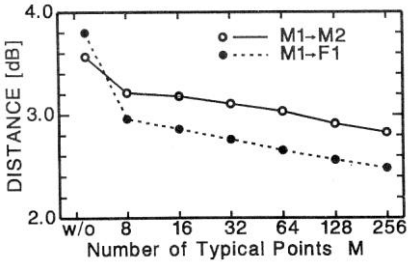
Dimana, $i(r), j(r)$ adalah nomor frame untuk deret spectral target dan transformasi di titik r pada DP path, sedangkan L adalah jumlah total kisi-kisi point DP path.

Tabel 1. Kondisi analisis

No.	Item	Nilai
1	Fundamental Frekwensi	10 kHz
2	Low-pass Filter	4.2 kHz
3	Analisis Window	Haming window (10ms)
4	Analisis Perioda	5 ms
5	Preemphasis	1-0.98 z^{-1}
6	Jumlah Point FFT	1.024
7	Jumlah Orde PSE	31 orde



Gambar 4. Hasil evaluasi pemetaan supervisi



Gambar 5. Hubungan sampel point

Hasil dari evaluasi memakai spektral distance seperti ditunjukkan pada Gambar 4 di bawah ini. Dari Gambar 4, dapat kita lihat bahwa nilai minimum distance spektral untuk kedua kombinasi transformasi suara terjadi pada saat p=1.0.

Dengan menetapkan p=1.0 maka hubungan antara spektral dan jumlah titik sampel M, dapat kita lihat seperti pada Gambar 5. Dari Gambar 5, dapat kita lihat bahwa dengan penggandaan jumlah titik sampel, maka distance spectral akan menurun sebesar 0.1dB untuk kedua kombinasi transformasi suara. Idealnya penentuan jumlah titik sampel disesuaikan dengan jumlah sampel training, akan tetapi meskipun jumlah sampel training dikurangi menjadi 25 kata, dan range M antara 1 – 256, distance spectral tetap mengalami penurunan secara rata,seperti pada gambar 5, maka untuk pemetaan supervisi ditetapkan p=1.0 dan M= 256

5.2 Evaluasi pemetaan unsupervisi

Sama seperti evaluasi untuk pemetaan spektral pada kondisi supervisi, tujuan dari evaluasi ini juga untuk melihat tingkat akurasi pemetaan spektral.

Kondisi database suara yang digunakan pada evaluasi ini, ditunjukkan pada Tabel 2 di bawah, sedangkan kondisi analisisnya sama seperti yang ditunjukkan pada Tabel 1 di atas.

Untuk parameter dan kondisi transformasi yang digunakan dalam evaluasi ini, ditunjukkan pada Tabel 3, sedangkan kombinasi transformasi suara pembicara, seperti yang ditunjukkan pada Tabel 4.

Tabel 2. *Source* Suara

No	Item	Nilai
1	Data Suara	28 Kata
2	Sampel Training	25 Kata
3	Speaker	Laki-laki 7 orang (m_1-m_7)
		Perempuan 4 orang (f_1 – f_4)

Tabel 3. Kondisi transformasi unsupervisi

No	Item	Nilai
1	Codebooks Size	1024
2	Jumlah model point M	32
3	Parameter interpolasi	1.0
4	Fuzziness	1.5
5	Limit Normal	1.0

Tabel 4. Kombinasi speaker transformasi suara

No	Transformasi suara	Kombinasi speaker
1	Laki-laki→Laki-laki m→ m	m_1→m_5 m_3→m_1 m_2→m_6 m_2→m_5
2	Laki-laki→Prmpuan m→ f	m_3→f_4 m_4→f_2 m_7→f_4 m_1→f_2

Evaluasi tingkat akurasi pemetaan transformasi deret spectral ke deret spectral target pada kondisi unsupervisi, juga menggunakan persamaan (13) spectral distance dan hasilnya seperti ditunjukkan pada Tabel 5 di bawah ini.

Tabel 5 menunjukkan hasil rata-rata spectral distance sebelum dan sesudah transformasi pemetaan, untuk tiap-tiap kombinasi transformasi suara pembicara. Selain itu, ditunjukkan juga spektral distance variasi pengucapan satu kata yang sama oleh referensi suara pembicara sebagai pembanding tingkat akurasi pemetaan spektral.

Dari Tabel 5 ini, dapat kita lihat bahwa, rata-rata spectral distance untuk transformasi suara laki-laki ke suara laki-laki, mengalami penurunan sekitar 60%, bila dibandingkan dengan rata-rata spektral distance sebelum

pemetaan, yaitu 2.3dB pada kondisi supervisi dan 2.2dB pada kondisi unsupervisi, serta sangat dekat dengan spektral distance referensi suara pembicara sebesar 2.1dB.

Sedangkan untuk transformasi suara laki-laki ke suara perempuan, rata-rata spectral distance juga mengalami penurunan dibandingkan dengan rata-rata spectral distance sebelum pemetaan, yaitu 2.4dB pada kondisi supervisi dan 2.3dB pada kondisi unsupervisi. Akan tetapi masih lebih besar dari spektral distance referensi suara pembicara sebesar 2.1dB.

Dari hasil ini juga terlihat, bahwa tingkat akurasi transformasi pemetaan pada kondisi unsupervisi, masih lebih rendah dibandingkan dengan kondisi supervisi. Hal ini dikarenakan masih terdapat transformasi input spektral itu sendiri, sehingga saat pendeskripsian kode spektral dengan pemetaan fuzzy terjadi sedikit kesalahan.

Tabel 5. Percobaan spektral distance suara (dB)

No	Transformasi suara	Sebelum transformasi	Supervised	Unsupervised	Variasi Pengucapan
1	$m_1 \rightarrow m_5$	4.2	2.2	2.2	2.2
2	$m_2 \rightarrow m_6$	4.1	2.5	2.4	2.2
3	$m_2 \rightarrow m_5$	4.2	2.3	2.2	2.2
4	$m_3 \rightarrow m_1$	2.8	2.1	2.1	1.9
	Rata - rata	3.7	2.3	2.2	2.1
5	$m_3 \rightarrow f_4$	3.4	2.4	2.3	2.2
6	$m_7 \rightarrow f_4$	3.6	2.5	2.4	2.2
7	$m_1 \rightarrow f_2$	4.1	2.3	2.3	2.0
8	$m_4 \rightarrow f_2$	3.7	2.3	2.3	2.0
	Rata - rata	3.7	2.4	2.3	2.1

6. Evaluasi transformasi kualitas suara berdasarkan identifikasi pembicara

6.1 Urutan percobaan

Evaluasi ini dilakukan untuk melihat tingkat akurasi kualitas suara hasil transformasi, dengan mengidentifikasi pembicara berdasarkan kelompok jenis kelamin. Sama seperti pada evaluasi sebelumnya, pertama-tama kita menyiapkan 8 macam kombinasi transformasi suara pembicara, yang dikelompokkan menjadi 2 kelompok, seperti pada Tabel 5 di atas. Selain itu kita siapkan juga 2 buah kata, yang kita kelompokkan masing-masing dalam kelompok A satu kata dan B juga satu kata.

Kemudian, kombinasi transformasi suara pembicara kita buat berpasangan dengan 2 buah kata yang telah dikelompokkan dalam A dan B, seperti yang ditunjukkan pada Tabel 6 di bawah ini.

Agar dapat dihasilkan tingkat akurasi semaksimal mungkin, maka diperlukan referensi suara sebagai pembanding yang mendekati suara alami asalnya, dengan cara melakukuan sintesis ulang input suara pembicara.

Untuk keperluan analisa hasil re-sintesis suara input dan suara referensi tersebut maka, kombinasi kata transformasi suara laki-laki ditetapkan (4+4)x2 kata, dari 5 orang pembicara laki-laki (m_1,m_2,m_3,m_4,m_5) yang berumur antara 20~40 tahun, dan kombinasi kata transfrormasi suara perempuan ditetapkan (4+2)x2 kata, dari 4 orang pembicara perempuan yang berumur 20 tahunan, sehingga jumlah total ada 169 pasang sampel kata, yang dipakai dalam percobaan kali ini.

Sedangkan fundamental frequency untuk masing-masing pembicara ditunjukkan pada Tabel 7. Dari tabel ini, dapat dilihat bahwa tidak ada perbedaan yang mencolok antara fundamental frequency pembicara, kecuali m_3.

Percobaan dilakukan dengan cara, meminta untuk setiap pasangan masing-masing mengucapkan 2 buah kata dalam kelompok A dan B, dengan interval waktu 13 detik. Kemudian dilakukan penentuan hasil dari pengucapan suara pembicara tersebut, apakah "Sama" atau "Tidak Sama" dengan suara referensi.

Tabel 6. Item pasangan suara (AxB)

A	B	Jumlah pasangan
m	m	5x5=25
f	f	4x4=16
m m	m	(4+4) 2x5=80
m f	f	(4+2) 2x4=48

Tabel 7. Rata-rata fundamental frekwensi masing-masing pembicara (Hz)

No.	Speaker	F_0
1	m_1	116
2	m_2	126
3	m_3	163
4	m_5	122
5	m_6	112
6	f_1	232
7	f_2	212
8	f_3	217
9	f_4	222

6.2 Hasil dan analisa

Dari percobaan ini, terdapat dua macam rasio jawaban yang benar, yaitu ratio kemungkinan pasangan yang mengucapkan [Sama], kita lambangkan dengan Ps dan ratio kemungkinan pembicara lainnya yang hasil ucapannya [Tidak Sama], kita lambangkan dengan Pd, terhadap suara referensi.

Apabila nilai Ps semakin besar, maka dapat dikatakan, kualitas suara hasil transformasi semakin mendekati suara referensi, akan tetapi, apabila standar sura referensi ditingkatkan, maka Ps akan menjadi lebih kecil dan nilai Pd menjadi besar.

Oleh karena itu, untuk melakukan evaluasi dalam percobaan kali ini, kita terlebih dahulu menetapkan rasio identifikasi (Pc) rata-rata Ps dan Pd dengan asumsi suara hasil transformasi sebagai suara referensi. Rasio identifikasi suara pembicara untuk setiap hasil suara transformasi, ditunjukkan pada tabel 8. Sedangkan pada tabel 9, menunjukkan hasil rata-rata Ps dan Pd untuk setiap suara hasil transformasi dari 5 orang pembicara.

Dari hasil pada tabel 8, dapat dilihat bahwa rasio identifikasi suara hasil transformasi laki-laki → laki-laki pada kondisi supervisi, lebih rendah 3,0% – 13,7% dibandingkan dengan suara referensi. Sedangkan pada kondisi unsupervisi, rata-rata identifikasi suara hasil

transformasi sebesar 84,0%, dapat digolongkan dalam transformasi yang cukup baik.

Rasio rata-rata identifikasi suara hasil transformasi laki-laki → perempuan, pada pembicara f4 dan f2 masing-masing adalah 76,4% dan 63,1%. Apabila dibandingkan dengan transformasi laki-laki → laki-laki, ternyata lebih rendah.

Dari hasil yang ditunjukkan pada tabel 9, dapat dilihat juga rasio rata-rata suara hasil transformasi untuk laki-laki → laki-laki, antara kondisi supevisi dan unsupervisi, dapat mendekati nilai yang hampir sama. Akan tetapi rasio rata-rata suara hasil transformasi untuk laki-laki → perempuan (khusus pada pembicara f4), terlihat nilai Ps pada kata kelompok A lebih tinggi, sedangkan Ps pada kata kelompok B lebih rendah, tetapi masih dapat dikatakan ada dalam level berpeluang untuk dapat lebih tinggi. Ini menunjukkan metode pemetaan yang kami perkenalkan, masih mempunyai keterbatasan untuk transformasi antara laki-laki dan perempuan.

Tabel 8. Rasio identifikasi pembicara setiap transformasi suara (%)

Transformasi suara	Supervised	Unsupervis-ed	Suara target
m_1 m_5	86,4	82,2	89,4
m_2 m_6	78,3	80,2	92,0
m_2 m_5	80,3	83,5	89,4
m_3 m_1	90,2	90,0	96,5
Rata-rata	**83,8**	**84,0**	**92,6**
m_3 f_4	79,4	76,8	88,7
m_7 f_4	73,3	64,1	88,7
m_1 f_2	68,3	--	93,5
m_4 f_2	57,9	--	93,5
Rata-rata	**69,7**	**70,5**	**91,1**

Tabel 9. Rasio rata-rata Ps-Pd setiap jenis suara (%)

Kata	Transformasi suara	Supervised		Unsupervised	
		P_s	P_d	P_s	P_d
A	mm	81.5	86.9	82.1	88.7
	mf	65.0	78.0	48.8	82.5
B	mm	77.6	88.7	74.8	90.1

Dari hasil percobaan yang telah dilakukan, rasio identifikasi pembicara untuk setiap suara hasil transformasi, sebagian besar mendukung spektral distance, seperti yang ditunjukkan pada tabel 5.

Seperti untuk kelompok yang sejenis kecuali pembicara m3 pada Tabel 7, dikarenakan tidak terdapat perbedaan fundamental frekwensi yang mencolok, maka dapat dikatakan bahwa, hasil identifikasi merupakan refleksi dari selisih kualitas suara berdasarkan amplop spektral.

Sebagai alasan rendahnya tingkat akurasi kualitas suara hasil transformasi suara laki-laki dan suara perempuan, salah satunya dikarenakan interpolasi selisih formant frekwensi belum dapat dilakukan dengan tepat, dan dikarenakan besarnya selisih spektral menyebabkan sangat sulit untuk menentukan spektral pendukung yang tepat diantara pembicara.

7. Penutup

Dari hasil yang diperoleh, dapat dikatakan bahwa pada transformasi suara laki-laki→suara laki-laki dengan jelas didapat hasil transformasi suara yang baik, dengan rata-rata ratio identifikasi sebesar 84%. Lalu, dari hasil analisa metode transformasi unsupervisi, diperoleh pula tingkat akurasi transformasi yang sama bila pada kondisi suara laki-laki→suara laki-laki.

Akan tetapi, pada transformasi antara suara laki-laki→suara perempuan, meskipun dengan metode supervisi menghasilkan tingkat akurasi yang rendah dibandingkan dengan transformasi suara laki-laki→suara perempuan, kemungkinan besar ini dikarenakan masih perlunya perbaikan/modifikasi pada tingkat akurasi pemetaan spektral, antara speaker yang memiliki perbedaan yang mencolok.

Untuk selanjutnya, dikemudian hari disarankan adanya modifikasi dan evaluasi kembali, agar dapat direalisasikan transformasi berdasarkan jenis individu pembicara dengan tingkat akurasi yang lebih tinggi lagi.

8. Daftar Pustaka

[1] K.Y. Park and H. S. Kim, "Narrowband to wideband convesion of speech using GMM based transformation", *Proc.ICASSP*, pp. 1847-1850. 2000.

[2] A. Kain and M. W. Macon, "Design and evaluation of a voice conversion algorith based on spectral envelope appint and residual prediction", *Proc.ICASSP*, pp. 813-816. 2001.

[3] Athnasios Mouchtaris, Jan Van der Spiegel, Paul Mueller, "Non-Parallel training for voice conversion by maximum liklihood constrained adaptation", *Proc.ICASSP*, pp.I-1-I-4. 2004.

[4] A. Oppenheim and D. Johnson, "Computation of spectra with unequal resolution using the fast Fourier transform", *Proc. IEEE 59*, pp. 299-301. 1971.

[5] Y. Linde, A. Buzo and R.M. Gray, "An algorithm for vector quantizer design", *IEEE Trans. Commun. COM-28*, pp. 84-95. 1980.

[6] K. Shikano, K. Lee and R. Reddy, "Speaker adaptation through vector quantization", *Proc. ICASSP 86*, pp. 2643-2646. 1986

[7] Robert Schalkoff, *Pattern Recognition : Statistical, Structural an Neural Approaches*, John Wiley & Sons, Inc. 1992.