Jurnal Linguistik Komputasional
Indonesian Association for Computational Linguistics

JLK
Vol. 1, No. 1
Maret 2018

Penjernihan Derau pada Suara Kanal Tunggal dengan Pembelajaran Faktorisasi Matriks Non-negatif tanpa Pengawasan

T. Tirtadwipa Manunggal [#1], Oskar Riandi [#2], Ardhi Ma'arik[#3], Lalan Suryantoro [#4], Achmad S. Putera [#5], Izzul H. Al-Hakam [#6]

#PT. Bahasa Kinerja Utama
Jalan Haji Naman Kompleks Lingga Indah No. 55 Bintara Jaya, Bekasi, Indonesia 13450

[1]tirta@bahasakita.co.id
[2]oskar@bahasakita.co.id
[3]ardi@bahasakita.co.id
[4]lalan@bahasakita.co.id
[5]satria@bahasakita.co.id
[6]izzul@bahasakita.co.id

Abstract*— This article examines an approach of denoising method on single channel using Non-negative Matrix Factorization (NMF) on unsupervised-learning scheme. This technique utilizes the property of NMF which unravels spectrogram matrices of noise-interfered speech and noise itself into their building-block vector. As extension for NMF, Wiener filter is applied in the end of steps. This method is designated to run in low latency system, hence preparing certain noise model for particular condition beforehand is impractical. Thus the noise model is taken automatically from the unvoiced part of noise-interfered speech. The contribution achieved in this research is the kind of NMF learning using linear and non-linear constraint which is done without explicitly providing noise models. Therefore the denoising process could be undergone flexibly in any noise condition.

***Keywords*— denoising, NMF, unsupervised learning**

Abstrak— Artikel ini mengulas pendekatan metode penjernihan derau pada suara kanal tunggal menggunakan Faktorisasi Matriks Non-negatif (NMF) dengan pembelajaran tanpa pengawasan. Teknik ini memanfaatkan sifat NMF yang mengurai matriks spektrogram suara terganggu derau dan suara derau itu sendiri menjadi komponen vektor penyusunnya. Sebagai penunjang NMF, filter Wiener diterapkan pada akhir tahapan. Penjernihan ini digunakan untuk sistem dengan latensi rendah, sehingga menyediakan model derau secara khusus di awal proses secara terpisah menjadi tidak praktis. Maka dari itu model derau diambil langsung dari suara yang akan dijernihkan. Kontribusi yang dicapai dalam penelitian ini adalah jenis pembelajaran NMF dengan perbandingan konstrain linier dan non-linier yang dilakukan tanpa secara eksplisit menyediakan model derau, sehingga penjernihan dapat digunakan secara lebih fleksibel untuk setiap kondisi derau.

***Kata kunci*— denoising, NMF, unsupervised learning**

I. PENDAHULUAN

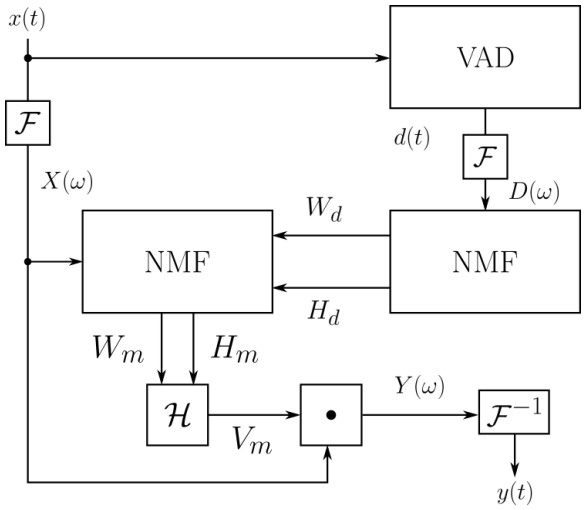
Interferensi pada suara ujaran (selanjutnya akan disebut dengan "suara") telah lama menjadi permasalahan baik di bidang telekomunikasi, penyiaran radio, penyiaran televisi, maupun pada mesin pengenalan suara otomatis atau *automatic speech recognition* (ASR). Gangguan tersebut sangat berpotensi menurunkan seberapa jelas informasi pada suara itu dapat dipahami. Sebagai gambaran, terdapatnya interferensi dapat secara langsung membuat akurasi mesin ASR merosot. Padahal jika keberadaan derau dapat ditekan, suatu sistem dapat bekerja dengan lebih baik. Dari sekian banyak jenis gangguan pada suara, salah satu jenis gangguan yang akan dibahas pada artikel ini adalah gangguan derau.

Derau pada suara bertumpang-tindih dengan suara asli. Hal ini akan makin terlihat pada domain frekuensi. Tumpang tindih tersebut mengurangi fokus pada ciri fonetik suara, sehingga boleh jadi spektrum derau lebih dominan dan suara menjadi tidak jelas terdengar. Untuk mengatasi permasalahan penjernihan derau ini, terdapat bermacam-macam cara yang telah diusulkan dan dilakukan. Secara garis besar, pendekatan-pendekatan yang ada terbagi menjadi dua yaitu *unsupervised learning denoising* (penjernihan derau dengan pembelajaran tanpa pengawasan) dan *supervised learning denoising* (penjernihan derau pembelajaran terawasi) [1].

Pendekatan penjernihan derau tanpa pengawasan mencakup beragam metode *spectral substraction* [2], *short-time spectral amplitude estimator* [3], tapis *impulse response*, penapis Wiener, penapis Kalman, dan lain-lain. Metode jenis ini melakukan tugas penjernihan derau tanpa pengetahuan sebelumnya mengenai model derau yang akan diredam. Tantangan utama dari pendekatan

Submitted 31-10-2017; Revised 27-12-2017; Accepted 26-02-2018

unsupervised learning ialah pengiraan kekuatan spektrum derau pada suara yang telah terinterferensi. Hal ini menjadi semakin sulit pada kasus ciri derau yang tidak stasioner.



Gambar 1 Gambaran umum metode penjernihan derau

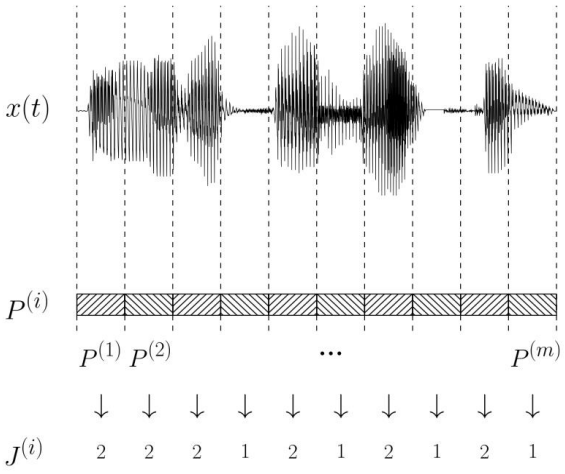
dalam masa pelatihan. Contoh metode penjernihan derau terawasi yaitu metode berbasis *Hidden Markov Model* (HMM) [4][5][6][7][8] maupun berbasis *codebook* [9] [10]. Teknik *denoising* terawasi akan unggul tergantung dengan apa yang diajarkan. Keunggulan ini dapat diperluas dengan menyediakan lebih banyak bahan ajar, namun tentu berbanding lurus dengan *computational cost* yang dibutuhkan.

Sebagai bentuk kompromi, terdapat penengah antara kedua pendekatan tersebut yang disebut dengan *semi-supervised denoising* berbasis *Non-negative Matrix Factorization* (NMF). Metode NMF pada peningkatan kualitas suara menggunakan matriks spectrogram sebagai basis. NMF mendekomposisi spectrogram menjadi matriks kolom dan vektor baris penyusunnya. Dengan sifat dekomposisi yang dimiliki NMF, membuat metode NMF banyak digunakan sebagai teknik separasi komponen suara.

Pada dasarnya, NMF bekerja dengan skema iteratif di mana pada setiap iterasi meminimalkan fungsi target. Sehingga tanpa tersedia informasi awal pun ia masih dapat bekerja. Akan tetapi, pemberian pengarahan pada NMF berupa matriks inisial dapat membawa pengaruh besar terhadap hasil akhir pengolahan suara, terlebih pada suara kanal tunggal. Pemisahan menggunakan NMF pada kanal tunggal cenderung lebih menantang daripada multi kanal karena terbatasnya informasi spektrum secara spasial.

Menilik sifat NMF, konsep penjernihan derau tetap membutuhkan informasi awal. Padahal mindset yang berlaku pada artikel ini adalah skema tanpa pengawasan. Perlu sebuah metode tambahan untuk memasok inisiasi NMF dengan model derau. Pada artikel ini, penyediaan model derau tersebut dilakukan secara otomatis memanfaatkan teknik *voice activity detection* (VAD) dengan asumsi derau bersifat kontinu walaupun tidak ada

Pada sisi yang lain, hasil dari metode dengan pengawasan tampak lebih superior. Kesiapan pendekatan jenis *supervised learning* untuk menghadapi derau terbangun dari informasi yang sebelumnya telah diberikan



Gambar 2 Ilustrasi proses k-means VAD

suara ujaran. Lazimnya VAD digunakan untuk mengambil bagian bersuara saja, namun dapat dipakai untuk menyisakan bagian tidak bersuara yang mengandung informasi model derau. Hasil yang digunakan ialah bagian *unvoiced* yang akan digunakan sebagai model derau.

TABEL I
RANGKUMAN NOTASI PENTING

$x(t)$	vektor suara dengan derau
$X(\omega)$	spektrum suara dengan derau
$d(t)$	vektor derau
$D(\omega)$	spektrum derau
$y(t)$	vektor suara hasil penjernihan
$Y(\omega)$	spektrum suara hasil pernjernihan
(s)	indeks penanda ujaran
(d)	indeks penanda derau
$P^{(m)}$	energi segmen suara
$\boldsymbol{K}$	banyak klaster k-means
μ	rataan klaster
$J^{(i)}$	fungsi objektif
ω	indeks frekuensi
τ	indeks spektro-temporal
$\mathcal{F}$	transformasi Fourier
h	fungsi window
V	matriks spektrogram $x(t)$
ϕ	fase spektrogram
W	matriks dictionary spektro-temporal
H	matriks model eksitasi
D_{KL}	divergensi Kullback-Leibler
$\mathcal{H}$	Wiener filter

Orientasi penyusunan artikel ini ialah pemaparan usulan langkah-langkah *unsupervised denoising*. Untuk membuat eksistensi derau lebih terukur, rekayasa adisi derau pada suara jernih juga dilakukan sebagai data uji. Derau yang ditambahkan adalah data nyata yang dapat ditemui pada kehidupan sehari-hari, bukan merupakan derau sintesis. Pada akhir bahasan, terdapat pula kesimpulan mengenai hasil penjernihan derau yang diperoleh.

II. Metode Penjernihan Derau

Pada bab ini akan dijelaskan mengenai metode VAD untuk menyisakan derau latar pada domain waktu dan metode NMF sebagai fitur kunci penjernihan derau. Secara umum alur kerja metode yang diusulkan ditunjukkan oleh diagram blok pada gambar 1. Adapun notasi penting terangkum pada tabel 1.

A. VAD Menggunakan K-means Clustering

Katakan vektor $x(t)$ selalu terbagi menjadi dua bagian yaitu bagian bersuara dan bagian diam. Bagian diam biasanya berupa jeda ucapan atau rehat sejenak. Tiap anggota $x(t)$ dikatakan sebuah bagian dari kelompok tertentu tergantung seberapa dekat kuantisasi nilai $x(t)$ dengan nilai rataan masing-masing kelompok. Bentuk kuantisasi $x(t)$ dapat berupa nilai amplitudo, amplitudo absolut, energi, dan lain-lain. Perbedaan mendasar antara bagian bersuara dan bagian yang tidak bersuara adalah energi suara itu sendiri, maka mari mengambil kuantisasi energi sebagai dasar pengelompokan. Perhitungan energi suara biasanya dilakukan secara berkelompok seperti yang tampak pada gambar 2. Hal ini disebabkan karena perhitungan energi satu per satu tiap anggota akan menjadi sangat fluktuatif, mengingat sinyal suara dalam domain waktu dapat dikatakan sinyal yang semi-acak. Energi suara pada domain waktu dihitung sebagai berikut,

$$P^{(m)} = 20 \cdot \log \left(\frac{\sum_{i=1}^{L} x(L \cdot m + i)^2}{I_{ref}} \right) \tag{1}$$

di mana $P^{(m)} \in \mathbb{R}^L$, L ialah lebar kelompok perhitungan energi, m adalah indeks untuk vektor energi, dan I_{ref} ialah referensi. $P^{(m)}$ merupakan vektor yang ingin dikelompokkan menjadi $\boldsymbol{K}$ klaster yang mana dalam hal ini $\boldsymbol{K} = 2$. Dengan premis tersebut maka akan terdapat dua *centroid* yaitu μ_1 untuk diam dan μ_2 untuk bersuara di mana $\min P \leq \mu_{\boldsymbol{K}} \leq \max P$ dan $\mu_1 < \mu_2$. Kedua *centroid* tersebut mula-mula bernilai acak sebab pada saat awal algoritma berjalan anggota masing-masing kelompok belum terdefinisi.

$$J^{(i)} = \arg\min_j \parallel P^{(i)} - \mu_j \parallel^2 \tag{2}$$

Algoritma *K-means* meminimalkan fungsi objektif dengan argumen J pada persamaan 2. Fungsi J tersebut menyortir tiap anggota $P^{(m)}$ menuju kelompok $j = 1$ atau $j = 2$. Dari pengejawantahan kelompok ini akan diperoleh pembaruan μ berikut,

$$\mu_j = \frac{\sum_{i=1}^{m} \mathbf{1}\left\{J^{(i)}\right\} \cdot P^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\left\{J^{(i)}\right\}} \tag{3}$$

di mana fungsi boolean $\mathbf{1}\left\{J^{(i)}\right\}$ bernilai,

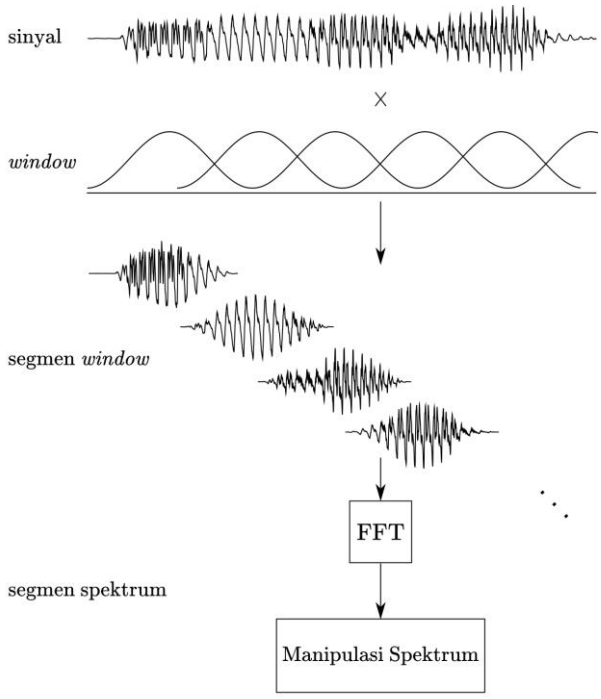
$$\mu_j = \frac{\sum_{i=1}^{m} \mathbf{1}\left\{J^{(i)}\right\} \cdot P^{(i)}}{\sum_{i=1}^{m} \mathbf{1}\left\{J^{(i)}\right\}} \tag{4}$$

Persamaan 2 dan 3 diulangi sebanyak N kali hingga fungsi J konvergen. Sehingga pada akhir masa iterasi, dengan asumsi distribusi normal, akan didapati μ_1 dan μ_2 yang mewakili nilai tengah tiap kelompok. Nilai μ_1 pada tataran implementasi dapat digeser mendekati μ_2 dengan suatu konstanta κ, sehingga $\mu_1 = \mu_1^* + \kappa(\mu_2 - \mu_1^*)$.

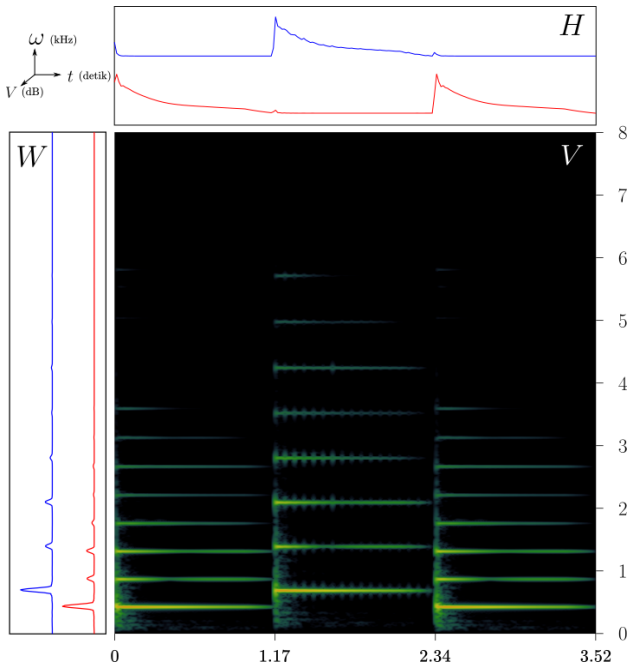
Kembali pada vektor energi P, pembagian kelompok tiap elemen P telah diperoleh. Untuk mencapai objektif penyisihan bagian kelompok diam, maka $x(t)$ dipilah berdasarkan indeks P yang terindikasi sebagai kelompok diam. Pemilahan disimpan dalam vektor yang dinotasikan sebagai $\boldsymbol{d} = \left[\boldsymbol{d}^*, x\left(t \mid J^{(i)} = 1\right)\right]$, di mana vektor $\boldsymbol{d}$ terus diimbuhi dengan $x(t)$ yang bersesuaian dengan kelompok diam.

Algoritma *Voice Activity Detection*

```
p ← 0
for i = 1, ··· , m do
    if J^(i) =: 1
        p ← p + 1
        if p ≥ threshold
            d ← [d*, x(t | J^(i) = 1)]
        end if
    else if J^(i) = 2
        p ← 0
    end if
end for
```



Gambar 3 Langkah-langkah Short Time Fourier Transform (Sethares, 2007)



Gambar 4 Contoh faktorisasi komponen pada kasus pemisahan nada A (445 Hz) dan F (707 Hz)

Sampai tahap ini, masih terdapat celah yaitu adanya kemungkinan bagian *unvoiced* dari suatu kata yang dikategorikan sebagai kelompok diam. Supaya tidak ada bagian kata yang terpenggal, maka vektor $\boldsymbol{d}$ perlu dipilah lebih lanjut dengan mempertimbangkan tetangga sebelum-sebelumnya. Metode pemilahan lanjut ini bekerja sesuai algoritma *Voice Activity Detection*. Pada akhirnya model suara derau diperoleh dengan persyaratan pada persamaan (5). Vektor $\boldsymbol{d}$ atau $d(t)$ merupakan elemen dari $x(t)$ yang akan digunakan sebagai model derau.

$$\boldsymbol{d} = \left[\boldsymbol{d}^{*},\ x\left(t \mid J^{(i)} = 1 \wedge\ p \geq \text{threshold}\right)\right] \tag{5}$$

B. Analisis Spektrum

Spektrum suara membawa informasi penting tentang komponen penyusun suara utama yaitu frekuensi. Perubahan dari domain waktu menuju domain frekuensi dilakukan dengan menggunakan transformasi Fourier. Pada sebuah sinyal kontinu, transformasi Fourier didefinisikan sebagai berikut [11],

$$X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t} dt \tag{6}$$

di mana $X(\omega) \in \mathbb{C}$ dan i ialah notasi bilangan imajiner.

Persamaan 6 akan menjadi berbeda pada implementasi komputer sebab sifat sinyal tidak lagi kontinu melainkan diskrit. Pada sinyal diskrit, persamaan transformasi Fourier didefinisikan ulang dengan [12],

$$X[\omega] = \sum_{k=0}^{N-1} x[k] e^{-i\omega k} \tag{7}$$

dengan $X[\omega] \in \mathbb{C}$ dan N sebagai lebar transformasi. Walaupun terdapat perbedaan mendasar mengenai kontinuitas, artikel ini akan menggunakan dua terminologi dan notasi kontinu-diskrit secara bergantian demi kemudahan. Penyebutan transformasi tersebut juga lebih lazim dengan *Discrete Fourier Transform* (DFT) maupun *Fast Fourier Transform* (FFT) yaitu implementasi transformasi secara komputasional.

Berangkat dari transformasi Fourier, metode analisis spektrum semakin berkembang, salah satunya ialah *Short Time Fourier Transform* (STFT). Ilustrasi STFT dapat dilihat pada gambar 3. Jika pada transformasi Fourier suara akan kehilangan informasi temporal, maka dengan STFT informasi spektral akan berdampingan dengan informasi waktu.

Ide dasar dari STFT adalah melokalisasi FFT pada suatu rentang yang terbatas [13]. Makna lokalisasi di sini adalah mencuplik sebagian kecil sinyal yang berdekatan dari sinyal yang panjang. Pencuplikan ini dilakukan secara *overlapping* terhadap pencuplikan sebelum atau setelahnya dan secara menyeluruh melingkupi sinyal utuh. Untuk mencapai *overlapping* yang baik, pada setiap cuplikan diterapkan fungsi pembobot h (atau disebut dengan *window*). Terdapat banyak macam *window* yang dapat digunakan, pada artikel ini dipilih salah satu fungsi window yaitu fungsi *window* Hanning (persamaaan 8).

$$h[k] = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi k}{N-1}\right)\right) \tag{8}$$

Dari uraian ini, SFFT dapat dinotasikan sebagai berikut,

$$X_{\tau}[\omega] = \sum_{k=0}^{N-1} x[k] h[k-\tau] e^{-i\omega k} \tag{9}$$

dengan τ sebagai indeks spektro-temporal.

Segala rekayasa spektrum dilangsungkan pada $X_\tau[\omega]$. Setelah rekayasa yang dimaksudkan tuntas, $X_\tau[\omega]$ dapat dikembalikan ke dalam domain waktu dengan mengulang langkah pada gambar 3 secara terbalik. Langkah pada gambar 3 dan kebalikannya disebut dengan metode *overlap-and-add*.

Sebagai catatan, agar proses invers dapat kembali ke dalam domain waktu secara sempurna tanpa menyisakan bagian imajiner, maka $X_\tau[\omega]$ harus bersifat *conjugate symmetric*. $X_\tau[\omega]$ harus memenuhi syarat berikut,

$$X_\tau[\omega] = X_\tau^*[-\omega] \tag{10}$$

Sebagai metode alternatif, manipulasi spektrum juga dapat dilakukan dengan spektrogram. Perbedaan spektrogram V pada persamaan 11 dan matriks STFT adalah spektrogram merepresentasikan magnitudo dari matriks spektro-temporal STFT. Dengan kata lain $V \in \mathbb{R}_+^{m \times n}$. Spektrogram V inilah yang akan digunakan pada proses NMF.

$$V = |X_\tau[\omega]|^2 \tag{11}$$

Sekalipun V telah kehilangan informasi fase pada persamaan 11, V harus tetap dapat diinvers atau paling tidak dapat kembali menjadi $X_\tau[\omega]$. Untuk memenuhi kebutuhan tersebut, sebelum melalui persamaan 11, informasi fase antara bilangan riil dan imajiner harus direservasi ke dalam matriks $\phi_\tau[\omega] = \angle X_\tau[\omega]$. Menggunakan $\phi_\tau[\omega]$,spektrogram V dapat dikembalikan ke dalam bentuk $X_\tau[\omega]$ dengan menerapkan persamaan 12.

$$X_\tau[\omega] = \sum_{k=0}^{N-1} V_\tau[k] e^{i\phi_\tau[k]} \tag{12}$$

C. Penjernihan Derau dengan NMF

Pada permasalahan faktorisasi, spektrogram V dianggap tersusun atas matriks $W \in \mathbb{R}^{m \times k}$ dan $H \in \mathbb{R}^{k \times n}$ atau dapat dituliskan dalam bentuk,

$$V \approx WH \tag{13}$$

di mana ditentukan kemudian tergantung kebutuhan.

Faktorisasi NMF diilustrasikan seperti pada gambar 4 yang memisahkan secara sederhana komponen nada A dan F. Apabila berbicara secara umum matriks W dan H dapat berupa apapun, namun bila spesifik pada matriks spektrogram, biasanya W merepresentasikan model frekuensi (sering disebut dengan *dictionary*) dan H merepresentasikan model eksitasi dengan indeks waktu.

Definisi tersebut berasal dari pemodelan umum suara ujaran pada domain waktu sebagai fungsi *filter*. Suara $s(t)$ dihasilkan dari konvolusi model pita suara $\delta(t)$ dan sinyal sumber eksitasi suara $\epsilon(t)$ [14].

$$s(t) = \delta(t) * \epsilon(t) \tag{14}$$

Sinyal eksitasi $\epsilon(t)$ sendiri dapat dinyatakan sebagai penjumlahan dari sebarang koefisien frekuensi yang bersesuaian dengan frekuensi fundamental $\overline{\omega}$ [15].

$$\epsilon(t) = \sum_{k=1}^{p} c_k e^{i\overline{\omega}(t)k} \tag{15}$$

Dalam suatu durasi yang singkat $t \in \tau$ sedemikian hingga $\delta(t) = \delta_\tau(t)$ dan $\overline{\omega}(t) = \overline{\omega}_\tau$. Variabel c_k merupakan sinyal pulsa segitiga yang diaproksimasi dengan $c_k = \mathrm{sinc}\left(\frac{k\overline{\omega}}{2}\right)$. Menggabungkan persamaan 14 dan 15 menghasilkan,

$$s_\tau(t) = \sum_{k=1}^{p} c_k \hat{\delta}_\tau(k\overline{\omega}_\tau) e^{i\overline{\omega}_\tau kt} \tag{16}$$

di mana $\hat{\delta}_\tau = \mathcal{F}(\delta_\tau)$. Jika persamaan 16 jika dinyatakan dalam STFT, notasi akan menjadi,

$$|\hat{s}_\tau(\omega)| \approx \sum_{k=1}^{p} c_k |\hat{\delta}_\tau(k\overline{\omega}_\tau)||\hat{h}(\omega - k\overline{\omega}_\tau)| \tag{17}$$

di mana $\hat{h} = \mathcal{F}(h)$ fundamental dan digeser berdasarkan frekuensi $\overline{\omega}_\tau$. Kemudian diambil asumsi respon frekuensi $|\hat{\delta}_\tau(\omega)|$ pada sebarang τ merupakan kombinasi dari suatu spektrum dengan pembobot non negatif $\eta_j(\tau) \geq 0$ sedemikian hingga $|\hat{\delta}_\tau\omega| = \sum_{j=1}^{K} \eta_j(\tau)|\hat{\delta}_j(\omega)|$ sehingga persamaan 17 dimodifikasi menjadi

$$|\hat{s}_\tau(\omega)| \approx \sum_{j=1}^{K} \eta_j(\tau) \sum_{k=1}^{p} c_k |\hat{\delta}_\tau(k\overline{\omega}_\tau)||\hat{h}(\omega - k\overline{\omega}_\tau)| \tag{18}$$

Frekuensi ω perlu ditentukan kombinasinya kemudian agar harmonik terhadap ω_τ. Anggap saja penentuan ini dibatasi pada satu himpunan spektrum $\overline{\Omega}_L$ di mana $\overline{\omega}_{min} \leq \overline{\Omega}_L \leq \overline{\omega}_{max}$. Dari gabungan $\eta_j(\tau)$ dan $\overline{\Omega}_L$ akan terdapat M atau $K \times L$ kombinasi spektrum. Berangkat dari sini, *time-activation* H terbangun atas η_j dan *dictionary* W terbangun atas vektor penyusun δ_j untuk $j = 1, 2, \ldots M$. δ_j merepresentasikan satu deret frekuensi harmonik. Berkenaan dengan persamaan 18, elemen W dan H didefinisikan sebagai,

$$\begin{cases} W_j = \Psi_j \delta_j \\ H_j = \eta_j \end{cases} \tag{19}$$

di mana Ψ_j adalah matriks berisi konstanta yang membentangkan δ_j. Komponen η dan δ harus diinisiasi di awal. Untuk kemudian menotasikan bagian suara ujaran, W_j dan H_j akan dituliskan sebagai $W_j^{(s)}$ dan $H_j^{(s)}$.

Model untuk derau akan sedikit berbeda dari model ujaran di atas. Dengan asumsi bahwa derau tersusun atas komponen statis $|\hat{d}_\tau(\omega)|$ yang didefinisikan sebagai

$$|\hat{d}_\tau(\omega)| \approx \sum_{k=1}^{r} |\hat{d}_k(\omega)|\hat{\delta}_k^{(d)}(\tau) \tag{20}$$

dengan $|\hat{d}_\tau(\omega)|$ diperoleh dari model derau pada persamaan 5. Seperti yang dinyatakan sebelumnya bahwa model spektrum dikalikan dengan pembobot non negatif $\eta_j(\tau) \geq 0$ sedemikian hingga $\hat{\delta}_k^{(d)}(\tau) = \sum_{j=1}^{K} \eta_j^{(d)}(\tau)|\hat{\delta}_{kj}$ sehingga persamaan 20 untuk derau menjadi

$$\left|\hat{d}_\tau(\omega)\right| \approx \sum_{j=1}^{L} \eta_j^{(d)}(\tau) \sum_{k=1}^{r} |\hat{d}_k(\omega)|\hat{\delta}_{kj} \quad (21)$$

dan komponen penyusun W dan H untuk derau dinyatakan sebagai

$$\begin{cases} W_j^{(d)} = \Psi_j^{(d)}\delta_j^{(d)} \\ H_j^{(d)} = \eta_j^{(d)} \end{cases} \quad (22)$$

di mana $\Psi_j^{(d)} \in \mathbb{R}_+^{m\times r}$ merupakan bentuk spektrum derau. Implementasi penjernihan derau persamaan-persamaan di atas dilakukan dengan mengadaptasi konsep [17] dan [18]. Nilai $\Psi_j^{(d)}$ disarikan dari spektrogram derau dari persamaan 5.

Formulasi NMF pada persamaan 13 difaktorisasi menjadi komponen W dan H. Saat iterasi faktorisasi berlangsung nilai divergensi antara V dan WH terus dipantau.

Dalam artikel ini divergensi yang digunakan adalah divergensi β [16]. Secara lebih spesifik, divergensi yang dimaksud adalah divergensi $\beta = 1$ atau disebut dengan divergensi Kullback-Leibler yang dirumuskan seperti pada persamaan 25.

$$D_{KL}(V|WH) = \sum_{i,\ j} V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \quad (23)$$

Tiap iterasi mengusahakan optimasi fungsi objektif $D_{KL}(V|WH) + \lambda|H|$. Fungsi objektif ini termasuk ke dalam permasalahan linier ($L1$). Dengan algoritma *multiplicative heuristic* [16], pembaruan nilai W dan H dapat diselesaikan dengan $\nabla(V|WH)$ persamaan 24 dengan sebagai pembagian antar elemen V dengan WH dan $\mathbf{1} \in \mathbb{R}^{\mathbf{k}\times\mathbf{n}}$.

$$\delta_j \leftarrow \delta_j \cdot \frac{\Psi_j^T \nabla(V|WH)\eta_j^T}{\Psi_j^n \mathbf{1}\eta_\mathbf{j}^\mathbf{T}}$$
$$H \leftarrow H \cdot \frac{W^T\nabla(V|WH)}{W^T\mathbf{1}+\lambda} \quad (24)$$

Peningkatan performa pembaruan WH dapat direkayasa lebih lanjut dengan menerapkan fungsi objektif yang non-linier ($L2$) [19] seperti fungsi objektif berikut $D_{KL}(V|WH) + \lambda|H| + \alpha\sum_j |\delta_j|^2$. Pada permasalahan berikut, pembaruan nilai W dan H dirumuskan seperti pada persamaan 25.

$$\delta_j \leftarrow \tilde{\delta}_j \cdot \frac{\mathbf{1}_j\tilde{\delta}_j^T\Psi_j^T\mathbf{1}\eta_j^T + \Psi_j^T\nabla(V|WH)\eta_j^T + \alpha\mathbf{1}_j\tilde{\delta}_j^T\tilde{\delta}_j}{\Psi_j^T\mathbf{1}\eta_j^T + \mathbf{1}_j\tilde{\delta}_j^T\Psi_j^T\nabla(V|WH)\eta_j^T + \alpha\tilde{\delta}_j}$$
$$H \leftarrow H \cdot \frac{W^T\nabla(V|WH)}{W^T\mathbf{1}+\lambda} \quad (25)$$

Pada awal masa pembelajaran, $W_j^{(d)}$ dan $W_j^{(s)}$ digabungkan menjadi satu sebagai panduan optimasi. Sedangkan matriks *time-activation* diinisasi secara acak (lihat persamaan 26).

$$\begin{cases} \mathbf{W} = \left[W_j^{(d)}, W_j^{(s)}\right] \\ \mathbf{H} = \left[H_{(s)}^T, H_{(d)}^T\right]^T \end{cases} \quad (26)$$

Penjernihan derau kemudian dilakukan menggunakan tapis Wiener pada domain frekuensi berdasarkan metode NMF yang telah dilakukan. Untuk sampai ke sana, penting untuk menilik konsep tapis Wiener. Persamaan 14 dimodifikasi dengan notasi lain agar sejalan dengan notasi umum pada teori tapis Wiener seperti berikut

$$s(t) = h(t) * y(t) \quad (27)$$

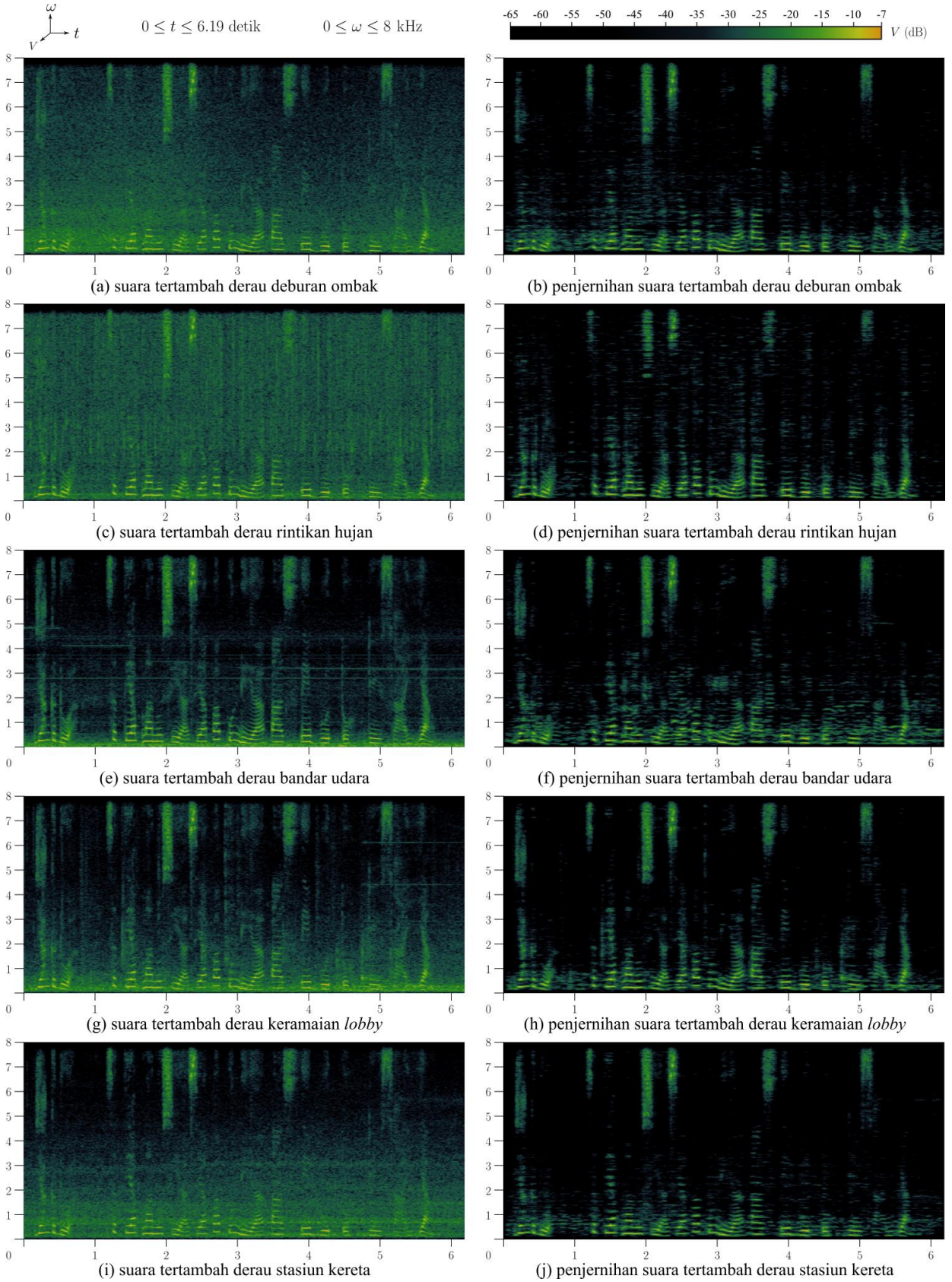
Pada domain frekuensi persamaan 27 dapat dituliskan sebagai berikut

$$S(\omega) = \mathcal{H}(\omega)Y(\omega) \quad (28)$$

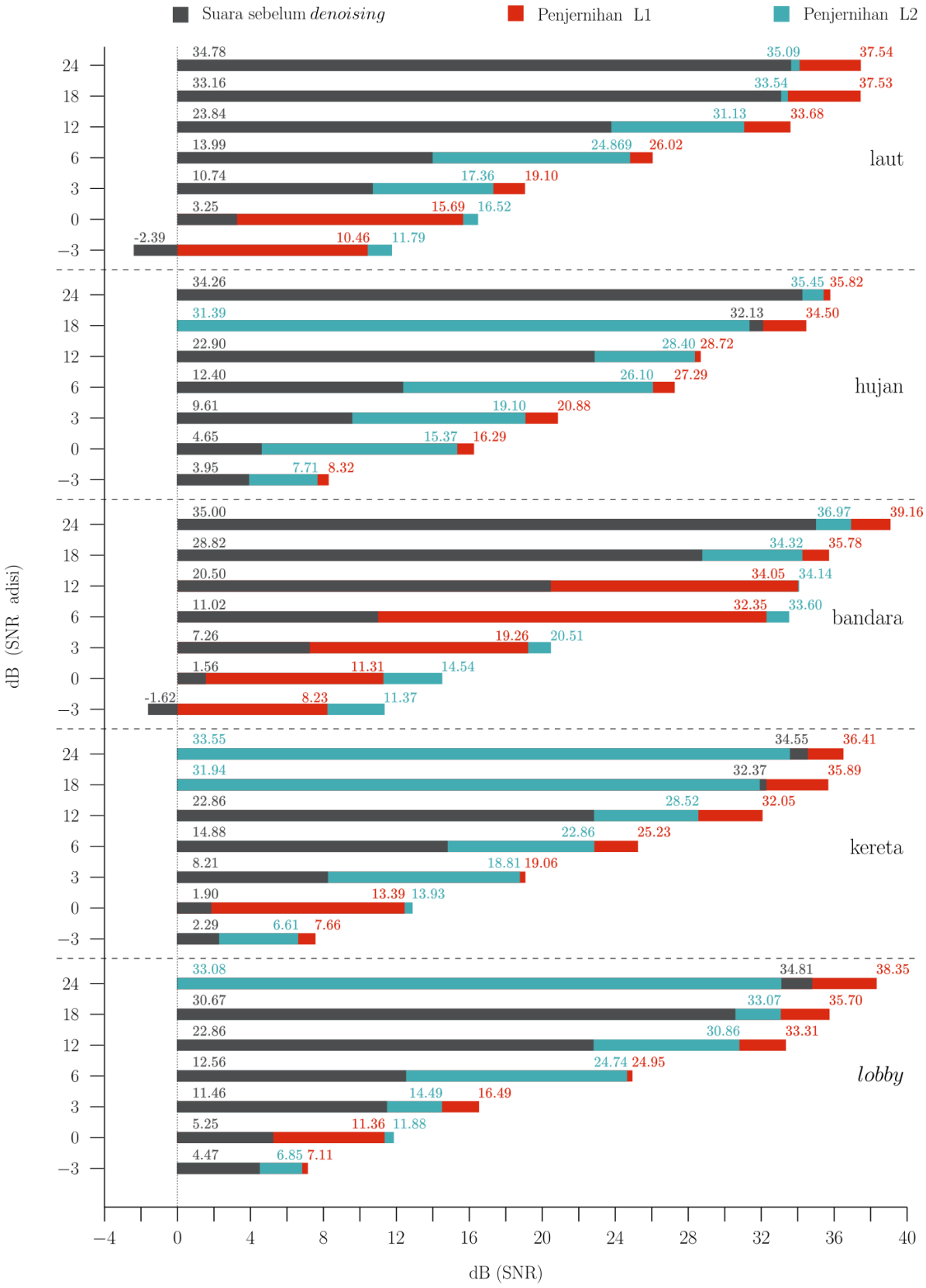
Persamaan 28 terdapat dua versi yaitu $\overline{S}(\omega)$ dan $\hat{S}(\omega)$, di mana simbol bar menandakan respon konvolusi yang diharapkan dan simbol topi menandakan respon hasil konvolusi. Perbedaan antara $\overline{S}(\omega)$ dan $\hat{S}(\omega)$ dapat dihitung dengan

$$\begin{aligned} E(\omega_k) =& \overline{S}(\omega_k) - \hat{S}(\omega_k) \\ =& \overline{S}(\omega_k) - \mathcal{H}(\omega_k)Y(\omega_k) \end{aligned} \quad (29)$$

dengan indeks ω_k melokalisasi frekuensi tertentu. Selanjutnya $\mathcal{H}(\omega_k)$ perlu dicari sedemikian hingga meminimalisir fungsi objektif *mean square error* J.



Gambar 5 Penjernihan ujaran "*... yang kedua setiap manusia siapapun dia pasti butuh disayang...*" dengan simulasi penambahan derau 6 dB



Gambar 6 Penjernihan derau adisi pada rentang -3 dB hingga 24 dB dengan membandingkan konstrain L1 (24) dan L2 (25)

Fungsi objektif pada [20] dinyatakan sebagai,

$$\begin{aligned} J =& E\left[|S(\omega_k)|^2\right] - H(\omega_k)P_{ys}(\omega_k) \\ & - H^*(\omega_k)P_{sy}(\omega_k) + |\mathcal{H}(\omega_k)|^2 P_{yy}\omega_k \end{aligned} \tag{30}$$

di mana $P_{yy}(\omega_k)$ adalah *power spectrum* dari $y(t)$ dan $P_{ys}(\omega_k)$ adalah cross-power spectrum dari $y(t)$ dan $s(t)$. Untuk mencari tapis optimum $\mathcal{H}(\omega_k)$, fungsi objektif J diturunkan terhadap $\mathcal{H}(\omega_k)$ dengan nilai turunan parsial nol dengan catatan bahwa $P_{ys}(\omega_k) = P^*_{sy}(\omega_k)$ (lihat persamaan 30).

$$\begin{aligned} \frac{\partial J}{\partial \mathcal{H}(\omega_k)} &= \mathcal{H}^*(\omega_k)P_{yy}(\omega_k) - P_{ys}(\omega_k) \\ &= \left[\mathcal{H}(\omega_k)P_{yy}(\omega_k) - P_{sy}(\omega_k)\right]^* \quad = 0 \end{aligned} \tag{31}$$

Sedemikian hingga diperoleh solusi tapis Wiener sebagai berikut,

$$\mathcal{H}(\omega_k) = \frac{P_{sy}(\omega_k)}{P_{yy}(\omega_k)} \tag{32}$$

Penjernihan derau dijalankan dengan menerapkan tapis $\mathcal{H}(\omega)$ pada sinyal $X(\omega)$ menghasilkan sinyal akhir $Y(\omega)$.

III. EKSPERIMEN

Untuk mendemonstrasikan performa penjernihan derau, digunakanlah suara yang diambil dari video pada situs MOOC IndonesiaX mata kuliah UT101 *Public Speaking* [21]. Data tersebut diambil di dalam studio sehingga suara ujaran murni tanpa derau latar. Suara ini ditambahkan dengan beberapa rekaman derau yaitu derau deburan ombak di laut, derau rintik hujan, derau jet pesawat di bandara, derau mesin kereta di stasiun, dan derau ujaran manusia pada *lobby*. Adisi derau ini dilakukan dengan parameter *signal-to-noise ratio* (SNR) sebesar -3 dB, 0 dB, 6 dB, 12 dB, 18 dB, dan 24 dB pada parameter konstrain $L1$ dan $L2$. Semakin kecil nilai SNR derau adisi, maka suara ujaran akan semakin sulit ditangkap. SNR sendiri dihitung sebagai berikut,

$$\mathrm{SNR} = 20 \cdot \log\left[\frac{\sum V_o^2}{\sum (V_o - V_d)^2}\right] \tag{33}$$

di mana V_o merupakan spektrogram suara original dan V_d merupakan suara yang telah mengalami adisi derau atapun suara hasil proses penjernihan derau.

Parameter implementasi algoritma diuraikan sebagai berikut. Frekuensi *sampling* yang dipilih adalah $fs = 16000$ Hz. Algoritma *Voice Activity Detection* diiterasi maksimum 10 kali dan menggunakan parameter $L = 32$, $\mathrm{threshold} = 10$, dan $\kappa = 0.5$. Spektrogram disusun dengan STFT selebar 128 ms dan overlap 75%. Penentuan lebar STFT terbilang cukup besar karena bila dikonversi ke jumlah sampel maka tiap *frame* STFT akan mengandung 2048 sampel. Hal ini mengingat algoritma ini didesain untuk berjalan pada *backend* sehingga *latency* proses tidak menjadi prioritas, namun resolusi spektrogram lebih diutamakan. Parameter penjernihan derau dengan NMF mengikuti hasil eksperimen pada [24] dengan asumsi parameter sebagai berikut : komponen derau $\eta_j^{(d)}$ sebanyak 16 kolom, $\varpi_{min} = 80$ Hz, $\varpi_{max} = 400$ Hz, $M = 132$, iterasi NMF sebanyak 25 kali, koefisien *sparsity* $\lambda_s = 0.2$ dan $\lambda_d = 0$, dan parameter regularisasi untuk $L2$ yaitu $\alpha = 10$.

Gambar 5 menunjukkan performa penjernihan derau dengan $L2$ pada beberapa kondisi derau. Dari gambar tersebut tampak karakteristik derau yang ditambahkan. Derau deburan ombak memiliki karakteristik yang mirip seperti *pink noise* karena dominan pada frekuensi rendah dan kemudian berangsur-angsur melemah pada frekuensi tinggi. Derau rintik hujan memiliki karakteristik yang mirip dengan *white noise* karena magnitudo frekuensi merata pada seluruh bagian. Derau bandara menganggu sebagian area frekuensi ujaran dan sangat tajam pada frekuensi menengah yaitu frekuensi mesin jet pesawat yang konsisten hingga akhir ujaran. Derau *lobby* mendemonstrasikan derau *cocktail party* [22] yang pada dasarnya derau berasal dari rentang frekuensi ujaran manusia, hanya saja kekuatannya lebih rendah dibanding ujaran utama. Dan derau pada stasiun kereta terletak pada frekuensi rendah dan menengah serta terdapat beberapa bagian yang dominan walaupun tidak tajam yaitu frekuensi suara mesin kereta.

Secara visual, penjernihan derau ombak menyisakan bagian yang bertumpang tindih dengan suara ujaran, sehingga masih menyisakan derau pada frekuensi rendah. Pada derau rintik hujan, derau telah diredam dengan merata walaupun masih terdapat derau yang terletak secara acak pada frekuensi tinggi. Pada derau bandara, dominansi frekuensi mesin jet pesawat dapat dihilangkan tidak bersisa. Pada derau *lobby* dan stasiun kereta menghasilkan spektrogram yang mirip dan menyisakan sebagian kecil derau pada frekuensi rendah.

Gambar 6 menyajikan perbandingan metode dengan konstrain $L1$ dan $L2$. Suara dengan adisi derau berwarna abu-abu, sedangkan suara yang telah diproses ditunjukkan dengan warna merah ($L1$) dan biru ($L2$). Adisi beberapa jenis derau menyebabkan menurunnya kualitas suara bahkan hingga memiliki nilai SNR negatif. Tampak pada gambar tersebut $L1$ banyak mengungguli $L2$ terutama pada adisi derau yang tidak terlalu parah (SNR adisi 6 dB hingga 24 dB). Adisi pada rentang ini dapat dibilang sebuah permasalahan *denoising* yang mudah karena masih memiliki nilai SNR yang cukup besar. Dengan kata lain, *intelligibility* suara masih baik. $L1$ sangat cocok untuk proses penjernihan secara umum dengan kompleksitas yang tidak terlalu rumit. Namun pada adisi derau hujan 18 dB, kereta 24 dB, kereta 18 dB, dan *lobby* 24 dB, konstrain $L2$ justru memperburuk kualitas suara. Hal ini tampak dengan nilai SNR $L2$ yang lebih rendah daripada suara yang belum diproses. Dari sini dapat diketahui bahwa $L2$ kurang cocok digunakan pada adisi derau dengan SNR tinggi.

Walaupun demikian, permasalahan *denoising* pada rentang SNR adisi 3 dB hingga -3 dB harus menjadi perhatian karena pada rentang ini, suara ujaran dan derau saling tumpah tindih sehingga menurunkan kejelasan

maksud suara ujaran. Pada beberapa jenis derau yaitu laut dan bandara, $L2$ menjernihkan derau lebih baik dari $L1$. Bahkan pada kasus derau bandara, $L2$ sangat signifikan mengungguli $L1$. Pada jenis derau tersebut $L2$ unggul dengan estimasi bagian derau dan ujaran yang kompleks. Pada jenis derau stasiun kereta, $L1$ dan $L2$ tidak secara tegas menunjukkan keunggulan antar metode. Pada bagian derau rintik air hujan, terdapat hal yang menarik yaitu $L1$ sama sekali memproses derau dengan lebih baik daripada $L2$. Keunggulan $L1$ disebabkan oleh jenis derau dengan karakteristik yang merata pada seluruh bagian spektrum, sehingga estimasi yang dilakukan $L1$ sudah cukup untuk melakukan kerja dengan baik.

IV. KESIMPULAN

Pada artikel ini telah dipaparkan metode penjernihan derau tanpa pengawasan yang fleksibel tanpa memberi informasi mengenai derau yang mana pada metode sebelumnya hal ini sangat dibutuhkan. Pembaruan yang diusulkan memanfaatkan metode *clustering* untuk mengambil model derau secara otonom dan NMF untuk memisahkan komponen ujaran dari derau, serta Wiener *filter* di akhir proses. Pemutakhiran komponen W dan H dilakukan secara iteratif dengan konstrain $L1$ dan $L2$. Konstrain $L1$ dan $L2$ tidak saling mengungguli satu sama lain, namun memiliki rentang atau karakteristik kerja tertentu untuk menghasilkan penjernihan yang baik. $L1$ cocok untuk bekerja pada jenis derau yang merata semisal *whitenoise* pada SNR adisi besar. Sedangkan $L2$ baik digunakan untuk jenis derau yang lebih spesifik pada SNR adisi yang rendah.

Demikian metode penjernihan derau tanpa pengawasan diusulkan. Dalam tataran implementasi masih banyak diambil asumsi, sehingga keandalan metode hanya teruji pada asumsi-asumsi yang dipilih. Sehingga, sebagai bentuk perbaikan pada penelitian lebih lanjut, perlu adanya kajian yang lebih dalam mengambil asumsi supaya permasalahan *denoising* dapat lebih efisien dan lebih baik memisahkan suara ujaran dan suara derau yang tidak diinginkan.

REFERENSI

[1] Mohammadiha, Nasser, Paris Smaragdis, and Arne Leijon. "Supervised and unsupervised speech enhancement using nonnegative matrix factorization." IEEE Transactions on Audio, Speech, and Language Processing 21.10 (2013): 2140-2151.

[2] Boll, Steven. "Suppression of acoustic noise in speech using spectral subtraction." IEEE Transactions on acoustics, speech, and signal processing 27.2 (1979): 113-120.

[3] Ephraim, Yariv, and David Malah. "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator." IEEE Transactions on Acoustics, Speech, and Signal Processing 32.6 (1984): 1109-1121.

[4] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," IEEE Trans. Signal Process., vol. 40, no. 4, pp. 725–735, Apr. 1992.

[5] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," IEEE Trans. Speech Audio Process., vol. 6, no. 5, pp. 445–455, Sep. 1998.

[6] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," IEEE Trans. Audio, Speech, and Language Process., vol. 15, no. 3, pp. 882–892, Mar. 2007.

[7] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-Gaussian priors," IEEE Signal Process. Letters, vol. 20, no. 3, pp. 253–256, Mar. 2013.

[8] H. Veisi and H. Sameti, "Speech enhancement using hidden Markov models in Mel-frequency domain," Speech Communication, vol. 55, no. 2, pp. 205–220, Feb. 2013.

[9] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," IEEE Trans. Audio, Speech, and Language Process., vol. 14, no. 1, pp. 163–176, Jan. 2006.

[10] T. V. Sreenivas and P. Kirnapure, "Codebook constrained Wiener filtering for speech enhancement," IEEE Trans. Speech Audio Process., vol. 4, no. 5, pp. 383–389, Sep. 1996.

[11] Bracewell, Ronald Newbold, and Ronald N. Bracewell. The Fourier transform and its applications. Vol. 31999. New York: McGraw-Hill, 1986.

[12] Harris, Fredric J. "On the use of windows for harmonic analysis with the discrete Fourier transform." Proceedings of the IEEE 66.1 (1978): 51-83.

[13] Welch, Peter. "The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms." IEEE Transactions on audio and electroacoustics 15.2 (1967): 70-73.

[14] Gold, Ben, Nelson Morgan, and Dan Ellis. Speech and audio signal processing: processing and perception of speech and music. John Wiley & Sons, 2011.

[15] McAulay, Robert, and Thomas Quatieri. "Speech analysis/synthesis based on a sinusoidal representation." IEEE Transactions on Acoustics, Speech, and Signal Processing 34.4 (1986): 744-754.

[16] Févotte, Cédric, and Jérôme Idier. "Algorithms for nonnegative matrix factorization with the β-divergence." Neural computation 23.9 (2011): 2421-2456.

[17] Schmidt, Mikkel N., Jan Larsen, and Fu-Tien Hsiao. "Wind noise reduction using non-negative sparse coding." Machine Learning for Signal Processing, 2007 IEEE Workshop on. IEEE, 2007.

[18] Cauchi, Benjamin, Stefan Goetze, and Simon Doclo. "Reduction of non-stationary noise for a robotic living assistant using sparse non-negative matrix factorization." Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments. Association for Computational Linguistics, 2012.

[19] Lyubimov, Nikolay, and Mikhail Kotov. "Non-negative matrix factorization with linear constraints for single-channel speech enhancement." arXiv preprint arXiv:1309.6047 (2013).

[20] Loizou, Philipos C. Speech enhancement: theory and practice. CRC press, 2013.

[21] Recht, Ben, et al. "Factoring nonnegative matrices with linear programs." Advances in Neural Information Processing Systems. 2012.

[22] Sri Sediyaningsih (Oktober 2017). UT101 Public Speaking. Diambil dari www.indonesiax.co.id