PENGENALAN PENGUCAP OTOMATIS UNTUK APLIKASI FORENSIK DI INDONESIA BERBASIS PEMODELAN I-VECTOR

Fian Adinata, Jocelyn Hartanto, R. Triyogo dan Miranti Indar Mandasari
Program Studi Teknik Fisika – Institut Teknologi Bandung
September – 2020

ABSTRAK

Sistem pengenalan pengucap adalah suatu proses teknologi yang dapat dilakukan untuk mengidentifikasi identitas pengucap dari suara ucapannya. Sistem ini dapat digunakan salah satunya untuk keperluan forensik. Di Indonesia, sistem ini secara aktif dipakai guna membantu proses verifikasi pengucap sebagai barang bukti di persidangan oleh Komisi Pemberantasan Korupsi (KPK), Kepolisian dan Kejaksaan. Saat ini, sistem yang umumnya digunakan adalah sistem pengenalan tergantung teks (text-dependent) yang memerlukan waktu lama dan campur tangan manusia. Oleh karena itu diperlukan suatu sistem yang dapat mempersingkat waktu analisis yang dibutuhkan dalam proses verifikasi namun dengan galat serendah mungkin. Salah satu solusi yang ada adalah dengan menggunakan sistem pengenalan pengucap otomatis tidak tergantung teks (text-independent).

Sistem yang dibuat dalam penelitian ini adalah sistem pengenalan pengucap otomatis dengan model Identity Vector (I-Vector). Sistem ini akan dilatih dan diuji dengan menggunakan basis data suara ucap berbahasa Indonesia yang diperoleh dari pengambilan data suara pada ruang semi-anechoic pada Laboratorium Akustik Adhiwijogo, Institut Teknologi Bandung. Dalam sistem ini data suara ucap akan diekstraksi fiturnya dengan menggunakan Mel Frequency Cepstral Coefficient (MFCC). Selain koefisien MFCC sebesar 19+1 dimensi, digunakan juga nilai delta MFCC dan delta-delta MFCC yang masing-masing memiliki 20 dimensi agar didapatkan informasi perubahan suara untuk melengkapinya. Data yang telah diekstraksi fiturnya akan dimodelkan dengan pemodelan I-Vector dengan menggunakan 32 komponen Gaussian dan 100 dimensi I-Vector. Selanjutnya dilakukan penilaian terhadap kemiripan sampel K dan UK dengan menggunakan perhitungan cosine distance.

Hasil penilaian itu digunakan untuk mengukur kemampuan sistem dalam mengenali sampel K dan UK berasal dari pengucap yang sama (target) atau tidak (non-target). Hasil ini dinormalisasi dengan 3 jenis normalisasi yaitu Zero Normalization (Z-norm), Test Normalization (T-norm), dan Zero-Test Normalization (ZT-norm). Digunakan data uji sebanyak 46 data suara ucap laki-laki dan 52 data suara ucap perempuan dengan data latih dari 20 data pertama untuk masing-masing jenis kelamin. Nilai EER terendah yang dapat dicapai oleh sistem ini adalah sebesar 3.50% yang berasal dari skenario wawancara perempuan yang telah dinormalisasi dengan T-norm dan ZT-norm.

Kata kunci: Pengenalan pengucap otomatis, Bahasa Indonesia, *I-Vector*, MFCC.

PENDAHULUAN

Sistem pengenalan pengucap atau speaker recognition merupakan suatu proses teknologi yang dapat dilakukan untuk mengidentifikasi identitas pengucap dari suara ucapannya. Sistem ini meniru cara kerja pendengaran manusia yang mengenali identitas seseorang dari ciri khas yang unik pada suaranya. Penggunaan sistem pengenalan pengucap salah satunya adalah pada bidang forensik untuk menjadikan bukti fisik berupa sampel rekaman suara ucap sebagai bukti yang sah di pengadilan.

Dalam forensik sistem ini menggunakan 2 jenis sampel yakni sampel *Known* (K) yang didapatkan selama proses penyidikan tersangka dan sampel *Unknown* (UK) yang merupakan bukti rekaman suara ucap. Di Indonesia yang telah menggunakan sistem pengenalan pengucap secara aktif guna membantu proses verifikasi pengucap sebagai barang bukti di persidangan adalah Komisi Pemberantasan Korupsi (KPK), Kepolisian dan Kejaksaan.

Saat ini, sistem pengenalan pengucap yang dipakai adalah sistem berbasis *text-dependent* yang artinya konten ucapan pada kedua sampel harus sama. Hal tersebut menjadikan sistem ini memakan banyak waktu karena dibutuhkan informasi tentang konten ucapan serta harus melalui proses *pairing* dan

manual tagging [1]. Untuk mengurangi waktu yang diperlukan, dikembangkanlah sistem berbasis *text-independent* yang tidak mempunyai batasan akan konten ucapan dari kedua sampel.

Saat ini di lingkungan Program Studi Teknik Fisika, Institut Teknologi Bandung telah dilakukan penelitian mengenai hal tersebut. Penelitian telah dilakukan pada tahun 2017 oleh Stefanus, I. dengan memanfaatkan *Gaussian Mixture Model* (GMM) serta tahun 2018 oleh Firmanto, A.D. yang menggabungkan sistem GMM dengan *Universal Background Model* (UBM) yang masing-masing mempunyai *Equal Error Rate* (EER) 6.45% dan 4.66% untuk laki-laki serta 12.9% dan 9.59% untuk perempuan [1][2].

Penelitian ini merupakan lanjutan dari kedua penelitian tersebut. Model *I-Vector* yang digunakan dalam penelitian ini dipilih karena dapat mencapai nilai EER sebesar 1.12% [3]. Untuk melakukan pembuatan sistem berbasis model *I-Vector* ini akan dilakukan juga pengambilan data suara ucap berbahasa Indonesia karena belum adanya basis data standar yang dapat langsung digunakan. Keluaran yang diharapkan dari penelitian ini adalah terbentuknya sistem pengenalan pengucap otomatis untuk aplikasi forensik yang memiliki galat lebih rendah dari sistem GMM dan GMM-UBM yang ada sebelumnya.

2 TEORI DASAR

2.1 Suara Sebagai Biometrik

Manusia secara umum serupa dengan manusia yang lainnya. Namun, disisi lain kita juga mempunyai beberapa hal yang berbeda antara satu dengan yang lain yang dapat kita gunakan untuk mengidentifikasi antar manusia yang disebut dengan biometrik. Biometrik adalah karakteristik fisik yang unik untuk setiap individu. Dalam aplikasinya, biometrik dapat digunakan dalam sebuah sistem dengan tujuan dua hal, yaitu identifikasi dan verifikasi. Identifikasi artinya membandingkan sebuah data biometrik dengan data yang terdapat di basis data guna mengenali siapakah pemilik data tersebut, sedangkan verifikasi artinya membandingkan data biometrik tersebut dengan data milik seseorang yang telah direkam sebelumnya dengan tujuan untuk memastikan identitas orang tersebut. Pengenalan biometrik adalah identifikasi otomatis terhadap masing-masing individu dengan berdasar kepada karakteristik anatomi dan perilaku seperti sidik jari, wajah, iris, dan suara. Pengenalan biometrik suara dilakukan dengan melihat *pitch*, nada, dan ritme bicara.

2.2 Ekstraksi Fitur

Dalam proses analisis pada sistem pengenalan pengucap otomatis, data suara ucap yang hendak diverifikasi harus diberikan perlakuan terlebih dahulu sebelum dibentuk model yang akan dibandingkan antara sampel *Known* (K) dan *Unknown* (UK). Perlakuan ini ditujukan untuk mengekstraksi fitur yang ada di dalam data suara ucap itu sendiri. Proses ekstraksi fitur ini dilakukan dengan 2 tahap, yaitu *Voice Activity Detection* dan *Mel Frequency Cepstral Coefficient* (MFCC).

Voice Activity Detection (VAD adalah proses pengidentifikasian secara otomatis terkhusus pada bagian kapan sinyal suara mengandung suara ucap dan yang tidak [4]. Pada prosesnya, sinyal suara mentah diproses menggunakan VAD untuk memisahkan sinyal suara dengan sinyal jeda. Setiap sinyal yang merepresentasikan kata diproses lebih lanjut sedangkan sinyal jeda akan dihapus. Untuk memisahkan kondisi sinyal suara ucap, digunakan perhitungan daya sinyal dan *zero crossing rate*. Daya sinyal melambangkan seberapa kuat sinyal dalam satuan waktu tertentu sedangkan *zero crossing rate* melambangkan seberapa sering sinyal suara melewati titik nol dalam satuan waktu tertentu.

Dalam proses untuk mengenali seseorang dari suaranya, sistem memerlukan petunjuk berupa fitur dari suara itu sendiri. Fitur ini dibagi menjadi 2, yakni fitur tingkat tinggi dan tingkat rendah. Fitur tingkat rendah seperti *pitch* menggambarkan struktur anatomi organ pengucap, sedangkan fitur tingkat tinggi menggambarkan aspek-aspek suara yang dipelajari seperti tata bahasa dan pemilihan kata. Pada sistem pengenalan pengucap otomatis yang digunakan adalah fitur tingkat rendah karena lebih mudah diekstraksi dan dikuantifikasi.

Mel Frequency Cepstral Coefficient (MFCC) adalah teknik ekstraksi fitur yang paling populer digunakan karena berdasarkan pada rentang frekuensi kritikal dari telinga manusia [5]. MFCC dilakukan dengan mengubah data suara ucap menjadi koefisien MFCC dengan cara mengubahnya ke domain frekuensi dan membobotkannya dengan pembobotan Mel, yang akan membobotkan nilai frekuensi menjadi lebih

subjektif dan sesuai dengan pendengaran manusia. Selain MFCC, digunakan juga nilai delta MFCC dan delta-delta MFCC yang berisikan informasi perubahan suara terhadap waktu agar didapatkan fitur suara ucap yang lebih lengkap dan dapat meningkatkan akurasinya.

2.3 Model

Fitur yang didapatkan dengan MFCC tidak dapat secara langsung digunakan untuk membedakan antar pengucap. Hal tersebut dikarenakan fitur hasil ekstraksi MFCC tidak dapat menggambarkan karakteristik fisik pengucap suara, dalam hal ini konfigurasi saluran vokal pengucap. Fitur MFCC hanya menghasilkan vektor fitur spektral yang tidak dapat digunakan sebagai pembanding antara pengucap yang satu dengan yang lain. Untuk membuat fitur-fitur tersebut dapat dibandingkan, diperlukan sebuah model yang dapat menggambarkan karakteristik dari pengucap itu sendiri.

2.3.1 Gaussian Mixture Model (GMM)

Salah satu model yang dapat digunakan adalah Gaussian Mixture Model (GMM). Dengan GMM vektor fitur direpresentasikan oleh jumlah bobot dari komponen fungsi kerapatan probabilitas Gauss dengan rata-rata dan kovarian tertentu.

2.3.2 Universal Background Model (UBM)

Universal Background Model (UBM) adalah model GMM yang dilatih dengan banyak dataset dari fitur pengucap untuk merepresentasikan distribusi fitur umum yang tidak bergantung kepada pengucapnya (speaker-independent). Untuk sistem dengan basis data yang sedikit, nilai UBM didapatkan dengan cara melatih model menggunakan vektor fitur diluar sampel K. Kumpulan vektor fitur itu dikumpulkan dan dimaksimalkan nilainya dengan menggunakan algoritma EM hingga didapatkan hasil λ_{UBM} yang merepresentasikan model pengucap dengan bobot, rata-rata, dan kovarian tertentu [6]. Cara tersebut menjadikan setiap sampel K mempunyai nilai UBM yang berbeda. UBM ini digunakan secara bersamaan dengan GMM. Gabungan GMM-UBM ini menjadikan perhitungan likelihood ratio lebih cepat dan meningkatkan performa jika dibandingkan dengan GMM yang dilatih secara terpisah [2].

2.3.3 I-Vector

Variabilitas channel pengambilan data, variabilitas sesi pelatihan dan pengujian menjadi salah satu faktor yang dapat mengurangi performa GMM-UBM. Salah satu solusi yang digunakan untuk mengatasi hal tersebut adalah *Joint Factor Analysis* (JFA). Pada JFA, data suara ucap direpresentasikan oleh sebuah supervector M yang terdiri atas kombinasi variabilitas pengucap dan channel. Model *I-Vector* adalah modifikasi dari JFA tersebut. Pada model ini, hanya ada 1 matriks variabilitas yang disebut ruang variabilitas total (*total variability space*).

Fitur dari MFCC dinormalisasi terlebih dahulu untuk mengurangi variabilitas *channel*. Kemudian dilakukan ekstraksi *I-Vector* dari *supervector* UBM dan matriks total variabilitas. Selanjutnya dilakukan kompensasi *channel* dan penilaian dengan *Probabilistic Linear Discriminant Analysis* (PLDA). Penilaian dilaksanakan dengan cara membandingkan *I-Vector* antara model untuk setiap pengucap dan suara ucap yang digunakan pada pengujian [7]. Karena model berbasis *I-Vector* ini merupakan pengembangan dari model GMM dan UBM, performanya dapat dikatakan lebih baik daripada kedua pendahulunya, dengan mencapai nilai *Equal Error Rate* (EER) sebesar 1,12% [3].

2.4 Skoring dengan *Cosine Distance*

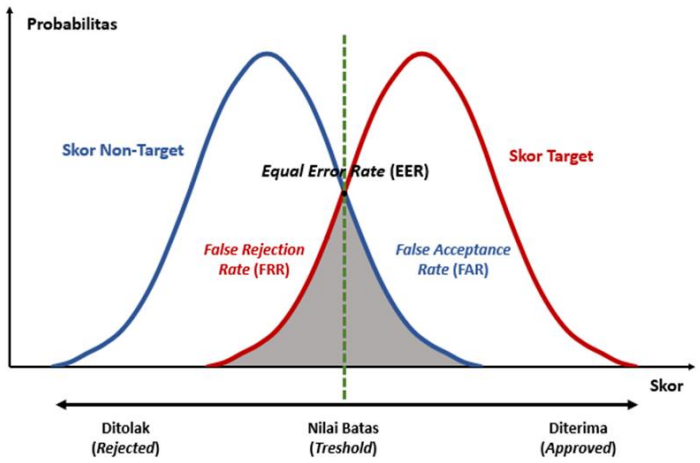
Perhitungan skor dengan *Cosine Distance* pada model *I-Vector* didasarkan pada perhitungan kemiripan dari vektor yang merepresentasikan fitur dari sampel K dan UK. Pada skoring ini, kemiripan vektor dihitung dari besar sudut antara kedua vektor yang sudah dinormalisasikan dengan menggunakan perkalian dot atau perkalian vektor dengan kosinus jarak diantara keduanya. Fungsi perhitungan dengan metode ini adalah sebagai berikut:

$$\text{Skor}\ (w_K, w_{UK}) = \frac{< w_K, w_{UK} >}{\| w_K \| \| w_{UK} \|} \quad (1)$$

dengan w_K menyatakan *I-Vector* dari sampel K dan w_{UK} menyatakan *I-Vector* dari sampel UK yang didapatkan dengan cara yang sama [3].

2.5 Normalisasi Skor dan Evaluasi dengan *Equal Error Rate* (EER)

Nilai yang didapatkan pada tahap skoring dibagi menjadi 2 jenis, skor target yang merupakan skor yang didapatkan jika sampel K dan UK berasal dari pengucap yang sama, serta skor non-target yakni skor apabila sampel K dan UK berasal dari pengucap yang berbeda. Ketika data kedua jenis skor itu dibuat menjadi grafik distribusi normal, akan terbentuk 2 grafik distribusi yang berpotongan.



Gambar 2.1 Grafik skor target dan skor non-target.

Dalam melakukan verifikasi, ditentukan sebuah nilai batas antara nilai target dan non-target, dengan nilai diatas nilai batas dianggap sebagai target dan nilai dibawahnya dianggap non-target. Hal tersebut mengakibatkan adanya skor-skor tertentu yang berada pada irisan kedua distribusi normal yang menjadi galat dari sistem. Galat ini dibagi menjadi 2 jenis, yakni *False Acceptance Rate* (FAR) yang berarti jumlah nilai yang seharusnya non-target namun dianggap sebagai target dan *False Rejection Rate* (FRR) yang artinya jumlah nilai target yang dianggap non-target oleh sistem. Penggeseran nilai batas ini akan mengakibatkan adanya *trade off* antara FAR dan FRR. *Equal Error Rate* (EER) adalah nilai yang menyatakan keadaan dimana nilai FAR dan FRR sama atau mendekati nilai yang sama.

Untuk memperbaiki nilai EER sistem, dapat dilakukan normalisasi terlebih dahulu pada skor mentah yang didapat. Normalisasi ini dilakukan untuk mengurangi variansi dari skor target dan non-target sehingga lebih mudah dibedakan antara keduanya. Cara ini digunakan untuk mengurangi EER sistem pada kondisi sistem yang tidak ideal. Dalam penelitian ini digunakan 3 jenis normalisasi skor yaitu *Zero Normalization* (Z-norm), *Test Normalization* (T-norm), dan *Zero-Test Normalization* (ZT-norm). Pada Z-norm normalisasi menggunakan estimasi rata-rata dan standar deviasi dari skor non-target untuk setiap sampel K, pada T norm menggunakan estimasi rata-rata dan standar deviasi dari skor non-target untuk setiap sampel UK, sedangkan pada ZT norm hasil T-norm kemudian dilanjutkan dengan normalisasi Z-norm.

3 RANCANGAN SISTEM

Sesuai dengan diagram alir yang terdapat pada Gambar 3.1, penelitian ini diawali dengan melakukan studi literatur mengenai sistem pengenalan pengucap otomatis khususnya yang menggunakan model *I-Vector.* Setelah itu, penelitian dimulai dengan adanya pengambilan data suara ucap Berbahasa Indonesia yang akan digunakan dalam pembuatan sistem pengenalan pengucap itu sendiri. Pengambilan data suara ucap ini dilakukan di Laboratorium Adhiwijogo, Institut Teknologi Bandung. Pengambilan data dilakukan dengan media 6 alat rekam yaitu *microphone* yang terhubung dengan laptop, *handphone* Lenovo A319, Sony ICD-PX333M *Digital Voice Recorder*, Zoom H2N *Handy Recorder*, Remax RP 1 *Digital Voice Recorder*, dan Olympus WS-852 *Digital Voice Recorder*. Setelah itu, data suara ucap yang terkumpul kemudian dikumpulkan menjadi 1 basis data yang dapat digunakan dalam penelitian ini.

Gambar 3.1 Metodologi Percobaan

Tabel 3.1. Perancangan basis data

Mode skenario	Skenario	Alat perekam
Text dependent	Huruf vokal (a, i, u, e, o)	Mikrofon dan 5 jenis alat perekam portable
	Angka (0 - 9)	
	Artikel (~300 kata)	
Text independent	Wawancara	
	Percakapan	

Data yang diambil kemudian disimpan dalam basis data dan dibedakan sesuai jenis kelamin, skenario pengambilan data, dan alat perekam yang digunakan. Selanjutnya data akan diproses terlebih dahulu untuk menghilangkan bagian data yang tidak diinginkan seperti derau selama perekaman dan disimpan dalam format *.wav*. Data ini akan digunakan sebagai data latih dan uji dalam pembuatan sistem.

Data uji yang digunakan dalam penelitian ini sebanyak 46 data suara ucap laki-laki dan 52 data perempuan, sedangkan data latih yang digunakan diambil dari 20 data pertama untuk masing-masing jenis kelamin. Data-data itu kemudian akan diekstraksi fiturnya dengan menggunakan VAD dan MFCC. Fitur data latih kemudian digunakan untuk melakukan pelatihan matriks total variabilitas dan UBM. Kedua matriks itu kemudian digunakan untuk mengekstraksi *I-Vector* dari data uji yang dibagi menjadi 2, dengan separuh bagian data awal dianggap sebagai sampel K dan separuh bagian data akhir dianggap sebagai sampel UK.

Sampel K dan UK yang telah diekstraksi *I-Vector*nya akan dinilai kemiripannya dengan menggunakan perhitungan *cosine distance.* Hasil perhitungan tersebut kemudian dinormalisasikan dengan menggunakan z-norm, t-norm, dan zt-norm. Semua skor yang didapatkan kemudian digunakan untuk menghitung EER sistem untuk dicari nilai terbaiknya.

4 HASIL DAN ANALISIS

Hasil skor dari setiap skenario termasuk skor hasil normalisasi dengan ketiga jenis normalisasi tersebut dihitung nilai EERnya dan ditabelkan. Selain itu, sebagai pembanding ditampilkan pula nilai EER pada sistem GMM-UBM yang dilakukan oleh Firmanto, A.D. pada tahun 2018 dengan skenario yang sama menggunakan 256 komponen Gaussian.

Tabel 4.1. Equal Error Rate sistem I-Vector untuk setiap scenario

Jenis Kelamin	Skenario	Raw	Z Norm	T Norm	ZT Norm
Laki-Laki	Percakapan	6.41%	4.23%	3.56%	4.81%
	Wawancara	7.57%	5.93%	5.73%	6.05%
Perempuan	Percakapan	12.78%	6.70%	6.48%	5.95%
	Wawancara	6.04%	3.80%	3.50%	3.50%

Tabel 4.2. Equal Error Rate sistem GMM-UBM untuk setiap skenario[2]

Jenis Kelamin	Skenario	Raw	Z Norm	T Norm	ZT Norm
Laki-Laki	Percakapan	41.30%	6.08%	41.60%	11.10%
	Wawancara	38.64%	4.66%	38.78%	9.83%
Perempuan	Percakapan	43.81%	19.11%	43.20%	23.08%
	Wawancara	43.39%	9.71%	43.20%	11.98%

Nilai EER dari sistem pengenalan pengucap dengan model *I-Vector* ini terdapat pada Tabel 4.1 diatas. Nilai EER terendah yang didapatkan sebesar 3.50% yakni pada hasil normalisasi skenario wawancara perempuan dengan menggunakan T-norm dan ZT-norm. Untuk skenario dengan pengucap laki-laki sendiri hasil EER terendah yang didapatkan adalah pada hasil skenario percakapan yang telah dinormalisasikan dengan menggunakan T-norm, dengan nilai EER sebesar 3.56%.

Dari nilai EER pada Tabel 4.1 dan 4.2 diatas juga dapat diketahui bahwa hasil perhitungan skor mentah dengan menggunakan sistem *I-Vector* ini jauh lebih baik dibandingkan hasil sistem GMM-UBM dalam skenario apapun. Walaupun kualitas sistem GMM-UBM telah diperbaiki dengan baik dengan menggunakan ketiga jenis normalisasi tersebut, nilai EER yang didapatkan menggunakan sistem *I-Vector* masih lebih rendah. Hal itu dengan pengecualian hasil EER Z-norm pada skenario wawancara laki-laki pada GMM-UBM yang lebih rendah sekitar 1.3% dari hasil dengan model *I-Vector*. Fakta ini menunjukkan bahwa model *I-Vector* yang digunakan dalam penelitian ini telah berhasil menurunkan galat dari sistem yang dikembangkan dalam penelitian sebelumnya yang berbasis model GMM-UBM.

5 KESIMPULAN

5.1 Kesimpulan

Pada penelitian ini telah berhasil dibuat sistem pengenalan pengucap otomatis berbasis pemodelan *I-Vector* dengan menggunakan 20 data latih dan parameter latihan berupa 32 komponen Gaussian dan 100 dimensi *I-Vector*. EER terendah dalam sistem yang dibuat mencapai 3.50% yakni pada hasil T-norm dan ZT-norm pada skenario wawancara perempuan. Sedangkan untuk laki-laki EER terendah yang didapatkan adalah 3.56% yang didapat dari skenario percakapan dengan dinormalisasikan menggunakan T-norm. Kedua nilai tersebut menunjukkan bahwa sistem ini telah berhasil menurunkan galat dari sistem sebelumnya yang berbasis model GMM-UBM.

6 DAFTAR PUSTAKA

[1] Stefanus, I. Perancangan sistem verifikasi otomatis untuk forensik suara ucap berbahasa Indonesia menggunakan Gaussian Mixture Model. Tesis Program Magister, Institut Teknologi Bandung. 2017.

[2] Firmanto, A.D. Pengembangan sistem verifikasi pengucap otomatis untuk kebutuhan forensic di Indonesia menggunakan model GMM-UBM. Tesis Program Magister, Institut Teknologi Bandung. 2018.

[3] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing, 19(4), 788–798. 2011. doi: 10.1109/tasl.2010.2064307.

[4] Rifqi, M. Evaluasi empiris dari gated unit recurrent neural network sebagai metode untuk voice activity detection. Skripsi Program Sarjana, Universitas Gadjah Mada. 2017.

[5] Hossan, M. A., Memon, S., & Gregory, M. A. A novel approach for MFCC feature extraction. *2010 4th International Conference on Signal Processing and Communication Systems*. 2010. doi: 10.1109/icspcs.2010.5709752

[6] Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, 10(1-3), 19–41. 2000. doi: 10.1006/dspr.1999.0361

[7] Tsujikawa, M., Nishikawa, T., & Matsui, T. I-vector-based speaker identification with extremely short utterances for both training and testing. *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*. 2017. doi: 10.1109/gcce.2017.8229389