PENGEMBANGAN MODEL DETEKSI DAN LOKALISASI BUNYI *NON-SPEECH* YANG TUMPANG TINDIH DENGAN MENAMBAHKAN TEKNIK PEMISAHAN BUNYI

DISERTASI

Karya tulis sebagai salah satu syarat
untuk memperoleh gelar Doktor dari
Institut Teknologi Bandung

Oleh
RANNY
NIM: 33217009
(Program Studi Doktor Teknik Elektro dan Informatika)

INSTITUT TEKNOLOGI BANDUNG
April 2024

ABSTRAK

PENGEMBANGAN MODEL DETEKSI DAN LOKALISASI BUNYI *NON-SPEECH* YANG TUMPANG TINDIH DENGAN MENAMBAHKAN TEKNIK PEMISAHAN BUNYI

Oleh
Ranny
NIM: 33217009
(Program Studi Teknik Elektro dan Informatika)

Teknik pengenalan bunyi *non-speech* telah banyak dikembangkan pada penelitian Teknik Informatika, begitu juga hasil penelitiannya telah banyak diimplementasikan pada beberapa aplikasi nyata. Namun, permasalahan terkait pengembangan teknik pengenalan bunyi *non-speech* masih terus ditelaah. Salah satu permasalahan yang dikembangkan adalah penggunaan bunyi *non-speech* yang tumpang tindih pada pengenalan bunyi. Bunyi tumpang tindih sendiri menjadi tantangan pada penelitian acuan karena sering kali menurunkan performa dari sistem pengenalan bunyi yang dibangun. Oleh karena itu, pada penelitian ini akan dikembangkan teknik yang mampu meningkatkan akurasi dari pengenalan bunyi dengan bunyi tumpang tindih. Teknik yang digunakan adalah teknik pemisahan bunyi tumpang tindih yaitu *Nonnegative Matrix Factorization* (NMF) dan *Time Frequency* (T-F) *Masking* yang kemudian dilakukan pengenalan bunyi menggunakan teknik pada *Machine Learning* antara lain, *Support Vector Machine* (SVM) dan *Artificial Neural Network* (ANN). Proses implementasi teknik dan pengujian menggunakan data *public* yang diaugmentasi untuk memperbanyak varian jenis bunyi tumpang tindihnya. Hasil eksperimen NMF dan SVM diukur dengan dimana nilai C menunjukkan tingkat *overfitting* dari model klasifikasi yang terbentuk. Semakin tinggi nilai C, maka semakin tingkat *overfitting*-nya. Rata-rata nilai C pada hasil pembentukan model klasifikasi adalah 4 dengan ukuran nilai C minimal 0 dan maksimal 20. Selain itu tingkat akurasi pengenalan juga diukur dalam persentase antara nilai positif dibagi hasil keseluruhan data. Rata-rata akurasi yang diperoleh adalah 83%. Berdasarkan hasil pengukuran teknik pemisahan NMF

dapat digunakan pada klasifikasi SVM. Pada eksperimen pemisahan T-F *masking* dan ANN diukur menggunakan nilai *F-score* dan *Error* rate-nya untuk kedua proses deteksi dan lokalisasi. Hasil rata-rata dari *F-score* untuk deteksi adalah 71,1% dan untuk lokalisasi adalah 81,5%. Sedangkan nilai *Error rate* untuk proses deteksi adalah 0,41 dan untuk lokalisasi adalah 12,5. Hasil dari pengujian dengan data bunyi tumpang tindih yang telah diaugmentasi menunjukkan peningkatan yang positif untuk teknik yang dikembangkan pada penelitian ini.

Kata kunci: bunyi tumpang tindih, pengenalan bunyi, pemisahan bunyi tumpang tindih.

ABSTRACT

DEVELOPMENT OF A MODEL FOR DETECTION AND LOCALIZATION OF OVERLAPPING NON-SPEECH SOUNDS BY ADDING SOUND SEPARATION TECHNIQUES

By

Ranny

NIM: 33217009

(Doctoral Program in Electro and Informatics)

Sound recognition techniques have been developed in Informatics Engineering research, and the results have been implemented in many applications. However, the problems related to the development of sound recognition techniques are still being studied. One of the problems developed is the use of overlapping sounds in sound recognition. Overlapping sounds themselves are a challenge in reference research because they often reduce the performance of the built sound recognition system. Therefore, in this study a technique will be developed that can improve the accuracy of sound recognition with overlapping sounds. The technique used is the overlapping sound separation technique, namely Nonnegative Matrix Factorization and Time Frequency Masking which is then carried out by sound recognition using techniques in Machine Learning, including Support Vector Machine and Artificial Neural Networks. The process of implementing techniques and testing uses augmented public data to increase the variants of overlapping sound types. The experimental outcomes of both NMF and SVM were assessed, wherein the parameter 'C' was used to signify the degree of overfitting within the resulting classification model. As the 'C' value increases, the level of overfitting also rises. In the classification model formation results, the mean 'C' value is 4, ranging from a minimum of 0 to a maximum of 20. Additionally, recognition accuracy was evaluated as a percentage derived from positive instances divided by the overall dataset outcomes, with an average accuracy of 83% achieved.Drawing from these measurement findings, the applicability of the NMF separation technique within SVM classification is evident. During the experimental phase, the disentanglement of T-F masking and ANN was evaluated through the utilization of F-score calculations and error rate analysis, spanning both detection and localization procedures. The average F-score for detection was computed at 71.1%, while for localization, it reached 81.5%. The error rate for detection was observed to be 0.41, and for localization, it was measured at 12.5. Encouragingly, testing with augmented overlapping sound data revealed a positive enhancement in the performance of the technique developed within the scope of this study.

Keywords: sound overlapping, sound recognition, overlapping sound separation.

PENGEMBANGAN MODEL DETEKSI DAN LOKALISASI BUNYI *NON-SPEECH* YANG TUMPANG TINDIH DENGAN MENAMBAHKAN TEKNIK PEMISAHAN BUNYI

Oleh
Ranny
NIM: 33217009
(Program Studi Doktor Elektro dan Informatika)

Institut Teknologi Bandung

Menyetujui
Tim Pembimbing

Tanggal 24 April 2024

Ketua

(Prof. Dr. Ir. Tati Latifah R. Mengko)

Anggota

(Dessi Puji Lestari ST, M.Eng., Ph.D.)

PEDOMAN PENGGUNAAN DISERTASI

Disertasi Doktor yang tidak dipublikasikan terdaftar dan tersedia di Perpustakaan Institut Teknologi Bandung, dan terbuka untuk umum dengan ketentuan bahwa hak cipta ada pada penulis dengan mengikuti aturan HaKI yang berlaku di Institut Teknologi Bandung. Referensi kepustakaan diperkenankan dicatat, tetapi pengutipan atau peringkasan hanya dapat dilakukan seizin penulis dan harus disertai dengan kaidah ilmiah untuk menyebutkan sumbernya.

Sitasi hasil penelitian Disertasi ini dapat ditulis dalam bahasa Indonesia sebagai berikut:

Ranny, R (2024): *Pengembangan Model Deteksi dan Lokalisasi Bunyi Non-Speech yang Tumpang Tindih dengan Menambahkan Teknik Pemisahan Bunyi*, Disertasi Program Doktor, Institut Teknologi Bandung.

dan dalam bahasa Inggris sebagai berikut:

Ranny, R (2024): *Development of A Model for Detection and Localization of Overlapping Non-Speech Sounds by Adding Sound Separation Techniques*, Doctoral Dissertation, Institut Teknologi Bandung.

Memperbanyak atau menerbitkan sebagian atau seluruh disertasi haruslah seizin Dekan Sekolah Pascasarjana, Institut Teknologi Bandung.

Dipersembahkan kepada orang tua, suami, adik kakak, mertua serta keluarga besarku tercinta yang senantiasa mendukung lahir dan batin.

KATA PENGANTAR

Puji Syukur penulis ucapkan kepada Tuhan Yang Maha Esa dan Maha Mengetahui, atas berkat dan tuntunan-Nya telah berhasil menyusun disertasi yang berjudul "Pengembangan Model Deteksi dan Lokalisasi Bunyi *Non-Speech* yang Tumpang Tindih dengan Menambahkan Teknik Pemisahan Bunyi," dengan baik dan lancar. Pada penyusunan disertasi ini penulis tak lepas dari bimbingan, dorongan, dan arahan dari berbagai pihak. Untuk itu secara khusus penulis ingin menyampaikan terima kasih yang sebesar-besarnya kepada Ibu Prof. Dr. Ir. Tati Latifah R. Mengko selaku ketua pembimbing yang telah sabar membimbing dan selalu mendorong untuk tidak menyerah dalam menyelesaikan disertasi ini. Penulis juga ingin mengucapkan rasa terima kasih yang besar kepada Ibu Dessi Puji Lestari ST, M.Eng., Ph.D. selaku anggota pembimbing, yang selalu meluangkan waktu untuk membimbing dengan sangat baik, serta mendorong penulis untuk melakukan penelitian dengan baik dan jujur. Penulis juga ingin mengucapkan terima kasih kepada Almarhum Bapak Prof. Dr. Ing. Ir. Iping Supriana Suwardi yang pada awal bimbingan adalah ketua pembimbing penulis yang telah memberikan peluang dan kesempatan bagi penulis untuk bergabung sebagai mahasiswa di Program Studi Doktor STEI ITB, semoga Almarhum diberikan tempat terbaik di sisi-Nya. Penulis juga mengucapkan terima kasih kepada Bapak Prof. Andriyan Bayu Suksmono, M.T, Ph.D, Ibu Dr. Eng. Ayu Purwarianti, S.T, M.T. dan Ibu Dr. Masayu Leylia Khodra, S.T, M.T. yang telah banyak memberikan masukan saat Seminar Kemajuan. Penulis juga mengucapkan terima kasih kepada para *reviewer* disertasi yaitu Bapak Prof. Dr. Ir. Bambang Riyanto Trilaksono selaku *reviewer* internal dan Ibu Prof. Lina, S.T., M.T, Ph.D selaku *reviewer* eksternal, yang telah membantu meningkatkan kualitas disertasi.

Kepada seluruh dosen dan staff di Program Studi Doktor STEI ITB, khususnya Ibu Nurhayati, A.Md. selaku TU Program Studi Doktor STEI ITB, penulis juga ucapkan terima kasih atas dorongan dan dukungannya selama menyelesaikan disertasi. Penulis juga ingin mengucapkan terima kasih kepada kawan-kawan: mahasiswa Doktor STEI ITB angkatan 2017, mahasiswa Lab GAIB STEI ITB, dan

mahasiswa Lab Robovis STEI ITB, yang setia menjadi teman diskusi dalam penulisan disertasi. Terima kasih juga penulis ucapkan kepada para dosen dan mahasiswa dari *School of Computer Science* BINUS@Bandung yang telah banyak memberi dorongan serta bantuan teknis selama penyusunan disertasi.

Penulis menyadari sepenuhnya bahwa karya tulis ilmiah ini sangat jauh dari kesempurnaan, sehingga saran dan kritik yang sifatnya membangun sangat dibutuhkan bagi kesempurnanaan disertasi ini. Demikian sedikit pengantar dari penulis, semoga disertasi ini dapat membawa manfaat terutama bagi penulis dan juga bagi pembaca.

Bandung, 24 April 2024

Ranny

DAFTAR ISI

DAFTAR LAMPIRAN

DAFTAR GAMBAR DAN ILUSTRASI

DAFTAR TABEL

DAFTAR SINGKATAN DAN LAMBANG

SINGKATAN	Nama	Pemakaian pertama kali pada halaman
MFCC	*Mel Frequency Cepstrum Coefficients*	2
SVM	*Support Vector Machine*	2
kNN	*k Nearest Neighbour*	2
LPCC	*Linear Prediction Cepstrum Coefficients*	2
GMM	*Gaussian Mixture Model*	2
HMM	*Hidden Markov Model*	2
ANN	*Artificial Neural Network*	5
Hz	*Hertz*	6
MMSE-STSA	*Minimum Mean Square Estimator Short-Time Amplitude Spectrum Estimator*	8
OCSVM	*One Class Support Vector Machine*	8
FFT	*Fast Fourier Transform*	9
DFT	*Discrete Fourier Transform*	10
STFT	*Short Time Frequency Transform*	10
DCT	*Discrete Cosine Transform*	11
Conv2D	*Convolutional Two Dimensional*	12
SED	*Sound Event Detection*	15
FOA	*first-order Ambisonics*	15
DOA	*direction-of-arrival*	15
DNN	*Deep Neural Network*	15
TDOA	*Time-difference-of-arrival*	15
SRP	*steered-response-power*	16
MUSIC	*multiple signal classificatio*	16
ESPRIT	*estimation of signal parameters via rotational invariance technique*	16
SNR	*signal-to-noise*	16
CNN	*Convolutional Neural Network*	16
RNN	*Recurrent Neural Network*	16
SELDnet	*Sound Event Localization and Detection*	17
ReLU	*Rectified Linier Unit*	18
GRU	*Gated Recurrent Unit*	18
LSTM	*Long Short-Term Memory*	18
FC	*Fully Connected*	18
NMF	*Nonenegative Matrix Factorization*	19
T-F	*Time-Frequency*	19
NUSSL	*Northwestern University Source Separation Library*	23
PCM	*array of pulse code*	23
ER	*error rate*	34
NIGENS	*Neural Information processing group General Sounds*	37

LAMBANG

h(t)	*Hamming window*	10
X(n,k)	Nilai STFT	10
H	*hop size* pada proses *Hamming window*	10
f	frekuensi	10

Bab I Pendahuluan

Pembahasan utama disertasi ini memusatkan perhatian pada salah satu aspek krusial dalam ilmu komputasi dan teknologi informasi, yaitu pengolahan bunyi. Bunyi merupakan elemen fundamental dalam interaksi manusia dengan lingkungan di sekitarnya. Dalam era digital yang semakin berkembang, kemampuan untuk memahami, menganalisis, dan memanipulasi data bunyi memiliki dampak yang besar dalam berbagai bidang seperti komunikasi, hiburan, kedokteran, dan kecerdasan buatan. Bab pendahuluan ini bertujuan untuk menguraikan latar belakang, tujuan, dan relevansi dari penelitian ini dalam konteks perkembangan teknologi pengolahan bunyi. Selain itu, bab ini juga akan memberikan gambaran umum tentang struktur dan metodologi penelitian yang akan dilakukan dalam disertasi ini untuk mendapatkan pemahaman yang mendalam dalam bidang pengolahan bunyi.

I.1 Latar Belakang

Pertumbuhan pesat dalam teknologi komunikasi dan multimedia telah membawa perubahan besar dalam cara manusia berinteraksi dengan informasi audio. Hal ini mengakibatkan timbulnya tantangan baru seiring dengan peningkatan kompleksitas data bunyi. Selain itu, pertumbuhan pesat dalam penggunaan perangkat elektronik seperti ponsel pintar, *speaker* pintar, dan asisten suara cerdas telah memperkuat pentingnya pengolahan bunyi dalam konteks kehidupan sehari-hari. Oleh karena itu, pemahaman mendalam terhadap metode-metode pengolahan bunyi menjadi semakin krusial untuk memenuhi kebutuhan masyarakat modern yang semakin tergantung pada teknologi berbasis suara.

Data bunyi sendiri dapat dikelompokan berdasarkan jenis sumber bunyinya. Sebagai contoh adalah bunyi yang dihasilkan manusia untuk berkomunikasi dikenal dengan istilah *speech* atau suara bicara manusia yang digunakan untuk komunikasi dan menunjukkan ekspresi. Terdapat juga bunyi *non-speech* atau bunyi yang bukan suara manusia, bunyi ini dikenal dengan istilah *environment sound* atau bunyi lingkungan. Jenis bunyi *non-speech* inilah yang digunakan pada penelitian ini, dan

selanjutnya hanya disebut sebagai bunyi. Jenis bunyi dilabelkan berdasarkan benda yang menjadi sumber bunyi, misal bunyi alarm dan bunyi mesin. Selain itu juga dapat dilabelkan berdasarkan kejadian atau peristiwa yang menyebabkan terjadinya bunyi, misal bunyi langkah kaki, bunyi pintu diketuk, bunyi gelas pecah, dan lain sebagainya. Penggunaan data bunyi ini akan diproses untuk berbagai tujuan pada penelitian tentang pengolahan bunyi.

Pada bidang pengolahan bunyi terdapat beberapa topik penelitian yang banyak dilakukan, antara lain: klasifikasi jenis bunyi, pengenalan atau identifikasi jenis bunyi, pengelompokkan jenis bunyi (HAPILABS, 2016; Kim dkk., 2004; Piczak, 2015; Sailor dkk, 2017). Terdapat juga penelitian lain yang berkembang yaitu tentang pemanfaatan bunyi dan suara sebagai alat interaksi manusia dengan mesin (Avenue dan Hill, n.d.; Baba dkk., 2004). Topik penelitian mendasar juga dikembangkan, yaitu pengembangan metode esktraksi ciri suara dan bunyi (Chia Ai dkk., 2012; Ganchev, 2005; Gilke dkk., 2012), metode klasifikasi dan *clustering* bunyi dan suara metode penghilangan *noise* atau gangguan pada data bunyi dan suara (Vacher dkk., 2010; Valin, 2007), terdapat juga penelitian tentang metode pengenalan suara itu sendiri (Avenue dan Hill, n.d.; Harma dkk., 2005; Vacher dkk., 2010). Selain itu, topik tentang metode pengambilan data suara, terkait dengan alat dan teknik yang digunakan untuk merekam data suara atau bunyi yang akan digunakan sebagai data eksperimen pada penelitian, dilakukan pada bidang pengolahan bunyi. Penelitian yang membahas tentang penggunaan *tools* dan *software* pendukung berbasis bunyi dan suara juga menjadi topik penelitian yang menarik (Istrate dkk., 2006; Maunder dkk., 2008; Shimada dkk., 2020).

Pada umumnya, penelitian pengolahan bunyi menggunakan kondisi eksperimen yang tidak menunjukkan kondisi nyata, yaitu data bunyi telah direkam satu per satu pada waktu yang berbeda, namun pada kondisi nyata terdapat beberapa jenis bunyi yang terjadi bersamaan dalam waktu yang sama. Hal inilah yang menjadi tantangan pada penelitian bunyi untuk bisa mengenali jenis bunyi yang terekam pada satu waktu yang sama. Umumnya penelitian terdahulu, banyak menggunakan bunyi tunggal dalam penelitiannya. Seperti pada penelitian oleh Karol J. Piczak (2015) yang mengumpulkan data bunyi lingkungan dan melakukan klasifikasi bunyi

lingkungan. Teknik ekstraksi ciri yang digunakan adalah *Mel Frequency Cepstrum Coefficients* (MFCC) dengan menggunakan klasifikasi *random forest*. Hasil eksperimen menunjukan tingkat klasifikasi 44.3% untuk *random forest*, 39.6% untuk *Support Vector Machine* (SVM) dan 32.2% untuk *k Nearest Neighbour* (kNN) (Piczak, 2015). Terdapat penelitian sejenis dengan data yang sama dilakukan oleh Sailor (2017) yang telah berhasil meningkatkan akurasi secara signifikan menggunakan metode *Convolutional Restricted Boltzmann Machine* untuk pelatihan *filterbank* dari data mentah dan metode klasifikasi dengan *Convolutional Neural Network*. Hasil akurasi terbaik mencapai lebih dari 80% (Sailor dkk., 2017).

Selain itu, terdapat aplikasi yang dirancang dengan kondisi lingkungan yang nyata, misal aplikasi *Smart Homes*, dimana lingkungan eksperimen yang digunakan adalah ruangan apartemen yang memiliki bunyi lebih dari satu sumber bunyi (Vacher dkk., 2010). Pada penelitian oleh Vacher (2010), salah satu eksperimen yang dilakukan adalah melakukan klasifikasi jenis bunyi yang terdiri dari lima jenis bunyi yaitu: bunyi tepuk tangan, dering telepon, bunyi cuci piring, benda jatuh dan teriakan. Teknik yang digunakan untuk ekstraksi ciri adalah *Linear Prediction Cepstrum Coefficients* (LPCC) dan membandingkan pembentukan meodel klasifiksi antara *Gaussian Mixture Model* (GMM) dengan *Hidden Markov Model* (HMM) yang pada akhirnya digunakan adalah GMM karena GMM lebih baik untuk kondisi dengan lebih tinggi tingkat *noise*-nya (Vacher dkk., 2010). Hasil eksperimen menunjukkan hanya dua data (bunyi telepon dan bunyi tepuk tangan) yang memiliki tingkat akurasi cukup baik, sisanya masih dengan tingkat akurasi yang rendah kurang dari 60%. Data bunyi yang digunakan sudah berupa data yang saling tumpang tindih, namun kondisi tumpang tindihnya dilakukan secara acak. Hasil segementasi bunyi yang dilakukan tidak menunjukkan hasil akurasi yang cukup baik.

Terdapat penelitian lain yang juga menggunakan data bunyi tumpang tindih yang dilakukan oleh Cakir (2017). Pada salah satu hasil penelitiannya menunjukkan akurasi deteksi jenis bunyi memiliki akurasi yang belum cukup baik (Çakir dkk., 2017). Eksperimen yang dilakukan salah satunya adalah mendeteksi 16 bunyi atau bunyi lingkungan yang tumpang tindih. Teknik ekstraksi ciri yang digunakan

adalah MFCC dan teknik deteksinya adalah *Convolutional Neural Network, Recurrent Neural Networks* dan kombinasi keduanya. Hasil akurasi yang diperoleh rata-rata masih di bawah 60%, dimana hasil ini menunjukkan tingkat akurasi yang belum cukup tinggi dan masih perlu dikembangkan teknik pemisahan bunyi tumpang tindihnya.

Terdapat juga penelitian yang berfokus pada deteksi dan lokalisasi bunyi dengan mengembangkan arsitektur multi proses. Deteksi bunyi berfokus pada jenis bunyi yang terdengar pada suatu waktu. Bunyi yang teridentifikasi dapat lebih dari satu jenis dalam satu waktu atau disebut dengan bunyi tumpang tindih. Sedangkan untuk lokalisasi bunyi berfokus pada identifikasi arah datang bunyi. Sumber bunyi yang bergerak membuat pola bunyi yang berbeda dengan bunyi yang statis. Bunyi yang bergerak akan diidentifikasi berdasarkan derajat *Azimuth* dan *Elevated*-nya. Kerangka kerja yang digunakan untuk deteksi dan lokalisasi bunyi adalah SELDnet (Adavanne dkk., 2019). Pada arsitektur Algoritma SELDnet ini telah diujikan menggunakan data tumpang tindih dan data tunggal. Hasil pengujiannya menunjukkan akurasi untuk data tumpang tindih masih kurang dari akruasi dengan data tunggalnya. Hal ini menunjukkan performa SELDnet untuk data tumpang tindih masih menjadi kekurangan dari SELDnet, ini menjadi peluang penelitian yang akan digali lebih mendalam.

Pada bidang pengolahan bunyi terdapat suatu topik penelitian yang menggunakan bunyi musik, yaitu klasifikasi jenis bunyi instrument pada lagu. Teknik yang digunakan pada penelitian musik ini adalah dengan menggunakan teknik pemisahan bunyi (Manilow dkk., 2018a). Terdapat beberapa teknik pemisahan bunyi yang telah dikembangkan. Berdasarkan penggunaan data latihnya teknik pemisahan bunyi ini dapat dibagi menjadi dua yaitu *blind separation*, contohnya *Nonnegative Matrix Factorization* (NMF) dan *nonblind separation*, contohnya *Time Frequency* (T-F) *masking.* Teknik NMF sendiri telah dikembangkan juga pada beberapa penelitian untuk deteksi bunyi, salah satunya pada penelitian oleh Bisot, dkk (Bisot dkk., 2017). Pada penelitian tersebut dilakukan deteksi bunyi menggunakan NMF sebagai model pengenalan bunyi yang tumpang tindih. Akurasi dari metode ini sebesar 49,5 persen, yang artinya masih menjadi peluang untuk

ditingkatkan performanya. Kedua teknik pemisahan NMF dan T-F *masking* akan digunakan pada eksperimen dari penelitian disertasi yang dilakukan. Pemisahan dengan NMF akan digunakan untuk klasifikasi bunyi menggunakan SVM dan pemisahan dengan T-F *masking* akan digunakan pada deteksi dan lokalisasi bunyi menggunakan SELDnet.

Selain dengan menambahkan teknik pemisahan bunyi, proses augmentasi data juga dilakukan pada penelitian ini. Jumlah varian dari bunyi tumpang tindih ditingkatkan pada proses augmentasi sehingga dapat meningkatkan variasi model latih. Proses augmentasi yang dilakukan adalah dengan menggabungkan beberapa jenis bunyi tunggal pada kelompok data bunyi tunggal sehingga menjadi bunyi yang tumpang tindih. Kedua kontribusi baik dengan pengembangan teknik maupun dengan augmentasi data diharapkan dapat meningkatkan hasil akurasi deteksi dan lokalisasi bunyi.

I.2 Masalah Penelitian

Penelitian yang dilakukan berfokus pada deteksi dan atau lokalisasi bunyi. Pada penelitian acuan, teknik deteksi dan lokalisasi menggunakan teknik dengan *Neural Network* yang langsung membentuk model untuk deteksi dan lokalisasi. Teknik *Neural Network* yang dikembangkan ini disebut dengan SELDnet. Hasil dari penggunaan model yang terbentuk ini diujikan pada data bunyi tunggal dan bunyi tumpang tindih. Namun hasilnya menunjukkan bahwa data tumpang tindih memiliki akurasi yang masih kurang dari 60%, jauh dibawah akurasi data tunggalnya yang telah mencapi 80%. Permasalahan tentang data tumpang tindih inilah yang akan diangkat pada penelitian ini.

I.3 Tujuan Penelitiann

Berdasarkan permasalahan yang telah dijabarkan, tujuan penelitian yang ingin dicapai adalah merancang teknik pengenalan bunyi lingkungan yang mampu meningkatkan akurasi pada data tumpang tumpang tindih. Teknik yang dikembangkan adalah dengan menambahkan teknik pemisahan bunyi untuk menghasilkan data tunggal dari bunyi tumpang tindihnya. Kemudian dilanjutkan dengan proses deteksi dan lokalisasi bunyi. Selain itu tujuan penelitian ini juga

meningkatkan akurasi dengan menggunakan data augmentasi dari proses augmentasi data.

I.4 Lingkup Penelitian

Deteksi dan lokalisasi bunyi akan dijadikan sebagai lingkup penelitian. Sistem deteksi dan lokalisasi bunyi yang dikembangkan akan diuji menggunakan data tunggal dan data tumpang tindih. Penelitian yang dilakukan akan menggunakan data yang telah tersedia, antara lain ESC-50 (Piczak, 2015) untuk kemudian dimodifikasi agar sesuai dengan kondisi nyata yaitu bunyi tumpang tindih atau *overlapping*. Selain data ESC-50, data TAU-NIGENS 2020 juga akan digunakan pada penelitian ini. Isi dari data TAU-NIGENS 2020 terdiri dari data tanpa tumpang tindih dan data bunyi dengan tumpang tindih. Pada penelitian ini dilakukan proses augmentasi dari data tunggal TAU-NIGENS 2020 untuk memperbanyak jumlah varian dari data tumpang tindih. Jumlah jenis bunyi untuk setiap tumpang tindihnya berjumlah dua buah jenis bunyi yang berbeda. Hal ini menjadi salah satu batasan eksperimen yang dilakukan.

I.5 Premis dan Hipotesa

Premis:

Penelitian tentang sistem pengenalan bunyi telah menghasilkan akurasi yang tinggi untuk bunyi yang tidak saling tumpang tindih. Sistem pengenalan bunyi yang telah dikembangkan yaitu SELDnet belum dapat menghasilkan akurasi yang tinggi pada kondisi bunyi yang saling tumpang tindih.

Hipotesa:

Pengembangan metode pemisahan untuk deteksi dan lokalisasi dengan bunyi yang saling tumpang tindih akan meningkatkan akurasi pada sistem pengenalan bunyi lingkungan. Penggunaan data hasil pemisahan dan data tunggal sebagai data latih dapat meningkatkan akurasi pada sistem pengenalan, deteksi dan lokalisasi bunyi. Data augmentasi untuk meningkatkan jumlah varian dari data tumpang tindih dapat mempengaruhi hasil akurasi sistem.

I.6 Kontribusi

Kontribusi yang diharapkan pada penelitian ini adalah terdiri dari dua bagian yaitu kontribusi pada pengembangan metode yang dapat dimanfaatkan pada kondisi bunyi yang saling tumpang tindih. Berdasarkan hasil penelitian yang sudah ada, jika dikelompokkan berdasarkan penggunaan data latihnya dapat dibagi menjadi dua yaitu data latih dengan data tunggal atau data tumpang tindih (Cheong Took, dkk., 2008; Gao, dkk., 2011; Han, dkk., 2015), serta kelompok kedua adalah data latih dengan data tunggal dan data tumpang tindih (Li, dkk., 2009; Mogi dan Kasai, 2012; Tian, dkk., 2017). Sedangkan pada penelitian ini akan dikembangkan pendekatan berbeda yaitu menggunakan data tunggal dan data hasil pemisahan bunyi tumpang tindih. Penggabungan antara teknik pemisahan bunyi dengan deteksi dan lokalisasi bunyi menghasilkan algortima baru yang dapat meningkatkan akurasi dari sistem deteksi dan lokalisasi bunyi tumpang tindih. Bagian kedua adalah proses augmentasi data yang dilakukan dapat meningkatkan jumlah varian data pada proses pengenalan bunyi, sehingga mempengaruhi hasil akurasi secara positif. Proses augmentasi ini dapat mengatasi keterbatasan pengumpulan data secara langsung yang dimana pengumpulan data langsung lebih memerlukan waktu dan biaya.

Jika disimpulkan terdapat dua poin utama kontribusi yang dilakukan, antara lain:

1. Menggabungkan teknik pemisahan bunyi pada model deteksi dan lokalisasi bunyi yang menggunakan *Neural Network* SELDnet.
2. Melakukan proses augmentasi data yang prosesnya berbeda dari proses augmentasi data pada TAU NIGENS 2020 yang mana pada penelitian ini melakukan proses agumentasi dengan mensegmentasi dan menggabungkan kembali hasil segementasi sehingga menjadi data baru. Proses augmentasi ini menghasilkan data dengan lebih cepat karena pada data TAU NIGENS 2020 data dibuat dengan cara memutar data NIGENS pada alat putar suara (*speaker*) untuk kemudian direkam kembali menggunakan alat rekam suara (*michrophone*) untuk menjadi data baru. Proses TAU NIGENS 2020 ini memerlukan waktu dan biaya yang lebih banyak dibandingkan proses augmentasi yang dilakukan pada penelitian ini.

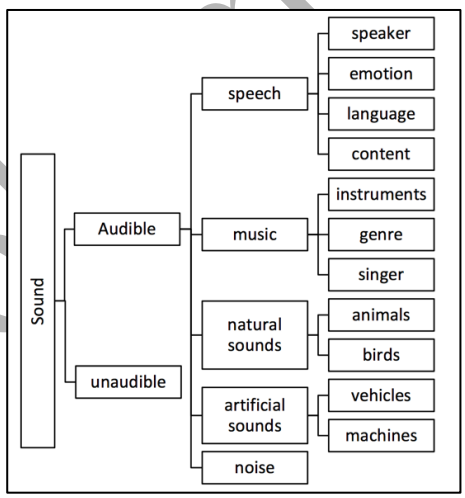
Bab II Tinjauan Pustaka

Pada bagian kajian teori ini akan dibahas beberapa teori dasar yang diperlukan untuk memahami tentang pemrosesan bunyi. Beberapa subtopik yang akan dibahas adalah dasar teori tentang bunyi: karakteristik bunyi dan bunyi lingkungan. Subbab berikutnya membahas tentang teori teknis pemrosesan bunyi secara digital yaitu: ekstraksi fitur bunyi, teknik pengenalan bunyi menggunakan *Artifical Neural Network* (ANN). Berkaitan dengan masalah penelitian, maka akan dibahas juga tentang teknik pemisahan bunyi tumpang tindih.

II.1 Karakteristik Bunyi

Pada lingkungan nyata, bunyi terdiri dari beberapa jenis, yaitu bunyi atau suara manusia, suara hewan, bunyi yang dihasilkan dari suatu kejadian atau benda mati, misal suara pintu diketuk, bunyi benda jatuh, bunyi klakson kendaraan, dan lain sebagainya. Bunyi merupakan gelombang yang dihasilkan akibat benda bergetar yang merambat pada zat tertentu (zat perambat) (Berg dkk., 1982; Pain, 2005b). Karakteristik gelombang bunyi yang dihasilkan dari sumber suara dapat berbeda satu dengan yang lain. Perbedaan karakteristik bunyi dihasilkan dari perbedaan sumber bunyi, media perambat bunyi, kecepatan bunyi, dan waktu (Berg dkk., 1982; Pain, 2005b). Perbedaan karakteristik gelombang bunyi inilah yang menyebabkan bunyi satu dengan yang lain bisa berbeda, meskipun dengan sumber bunyi yang sama. Walaupun bunyi terdengar berbeda, manusia memiliki kemampuan untuk membedakan bunyi tersebut, artinya bunyi memiliki ciri khusus yang digunakan untuk mengindetifikasi jenis bunyi. Identifikasi yang dilakukan bisa berupa sumber bunyi yang dihasilkan misal, bunyi suara manusia, sumber bunyinya dari manusia, bisa juga bunyi hewan artinya hewan sebagai sumber bunyinya atau bisa juga sumber bunyi dari benda mati, misal bunyi mesin kendaraan, artinya bunyi yang dihasilkan oleh mesin kendaraan. Selain itu bunyi juga dapat mengidentifikasi kegiatan yang menghasilkan bunyi, misal bunyi makan, bunyi memasak, bunyi berjalan dan lain sebagainya. Perbedaan bunyi ini dipengaruhi oleh frekuensi. Frekuensi bunyi yang dimaksud adalah jumlah gelombang yang dihasilkan dalam satu detik pada media rambat bunyi. Ukuran

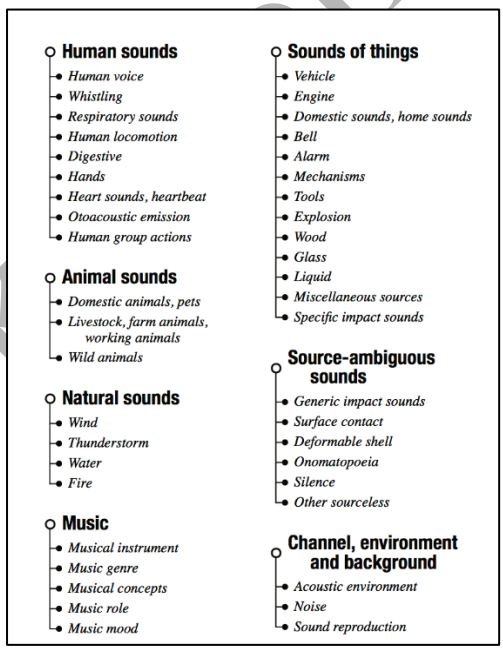
satuan dari frekuensi adalah *Hertz* (Hz). Manusia dapat mendengar bunyi dengan ukuran frekuensi 20 Hz – 20.000 Hz (Darji, 2017). Selain frekuensi faktor pembeda antar bunyi ada juga intensitas bunyi yang diukur dalam desibel (dB). Manusia dapat menangkap bunyi dengan ukuran 0 dB hingga 120 dB (Darji, 2017). Jika hanya menggunakan ukuran intensitas, maka bunyi dapat dikategorikan bunyi yang dapat nyaman terdengar dan tidak nyaman terdengar. Pada kenyataanya kategori bunyi lebih dari itu. Pengelompokan bunyi dapat dilihat pada Gambar II.1 (Darji, 2017). Kategori bunyi terdiri dari *audible* dan *unaudible*, atau bunyi yang terdengar dan bunyi yang tidak dapat didengar oleh manusia (diluar frekuensi tangkap manusia). Ketegori bunyi terdengar dibagi menjadi lima kelompok, yaitu *speech*, music, bunyi alam, bunyi buatan, *noise*. Masing-masing kelompok memiliki beberapa jenis bunyi lainnya. Kelompok *speech* memiliki empat jenis bunyi, antara lain *speaker*, emosi, bahasa, dan isi atau konten. Kelompok musikc memiliki tiga jenis, yaitu instrumen, genre, penyanyi atau *singer*. Kelompok *audible* berikutnya yaitu bunyi alam atau *natural sounds* yang memiliki dua jenis kelompok bunyi yaitu hewan dan burung. Kelompok bunyi buatan atau *artificial sounds* terbagi lagi menjadi dua yaitu kendaraan atau *vehicles* dan mesin atau *machines*.



Gambar II.1 Pengelompokkan Bunyi (Darji, 2017)

Selain pada pengelompokan bunyi oleh Darji (2017), terdapat juga ontologi bunyi yang dilakukan penelitian lain tentang jenis-jenis bunyi. Diagram ontologi dapat dilihat pada Gambar II.2 (Gemmeke dkk., 2017). Pada ontology ini terbagi menjadi enam kelompok jenis bunyi, yaitu *human sounds* atau bunyi yang berasal dari

manusia, jenis bunyi ini bukan hanya suara berbicara manusia, tapi juga bisa bunyi anggota tubuh lainnya misal bunyi jantung, paru-paru dan tepuk tangan. Kelompok kedua yaitu bunyi binatang atau *animls sounds*, jenis bunyi yang masuk pada kelompok ini adalah suara atau bunyi yang dihasilkan pada hewan peliharaan, hewan ternak serta hewan liar. Kelompok berikutnya adalah bunyi alam atau *natural sounds*, jenis bunyi yang masuk pada kelompok ini adalah bunyi-bunyi yang berasal dari alam seperti bunyi angin, badai, dan air mengalir. Bunyi benda adalah kelompok bunyi berikutnya, berbagai jenis bunyi benda seperti kendaraan, mesin, rumah, alaram, bel, peralatan pertukangan, gelas, kayu, cairan dan sumber bunyi benda lainnya, masuk pada kelompok ini. Kelompok berikutnya adalah kelompok sumber bunyi ambigu, jenis bunyi yang masuk kelompok ini adalah bunyi sesuatu jatuh, bunyi benda bertubrukan, dan bunyi bunyi ambigu lainnya. Kelompok bunyi terkahir adalah bunyi *channel*, lingkungan, dan latar, jenis bunyi yang masuk kelompok ini adalah bunyi akustik lingkungan, *noise*. Kelompok bunyi ini digunakan pada data set *Audio set* (Gemmeke dkk., 2017).



Gambar II.2 Diagram ontology bunyi (Gemmeke dkk., 2017)

Pada sistem pengenalan suara dan bunyi *non-speech*, *noise* menjadi penghambat pada peningkatan akurasi sistem. *Noise* yang dimaksud bisa dibedakan menjadi beberapa jenis, misal bunyi latar, bunyi gema, dan bunyi tumpang tindih. Namun,

keberadaan *noise* tidak dapat dihindari, mengingat lingkungan pengambilan data harus sealami mungkin atau sesuai dengan kondisi nyata saat sistem akan diimplementasikan. Sistem yang diterapkan pada suatu lingkungan yang tetap, misal: kantor atau tempat tinggal akan mempermudah sistem memprediksi *noise* yang akan terjadi. Namun, jika sistem digunakan pada berbagai lingkungan yang tidak dapat diprediksi, maka akan membuat *noise* menjadi sulit dihilangkan, misal: di stasiun, di dalam kendaraan umum, di jalan dan tempat lainnya yang memiliki banyak sumber suara dan bunyi. Terdapat beberapa metode untuk mengurangi atau menghilangkan *noise*. Pada penelitian yang dilakukan oleh Valin, dkk penghilangan *noise* pada sistem dilakukan dengan menggunakan *Echo Cancellation System* (Valin, 2007). Algoritma diimplementasikan pada SPEEX library dibawah lisensi GPL (Valin, 2007). Metode yang digunakan adalah *Minimum Mean Square Estimator Short-Time Amplitude Spectrum Estimator* (MMSE-STSA). Pada *noise* yang bersifat tumpang tindih, dapat dilakukan pemisahan bunyi.

II.2 Bunyi Lingkungan

Terdapat beberapa penelitian yang menggunakan data bunyi lingkungan. Salah satu tantangan dari penggunaan data bunyi lingkungan adalah variasi data yang sangat beragam, namun ketersediaan data masih terbatas. Salah satu implementasi dari hasil penelitian bunyi adalah deteksi suara jatuh, yaitu oleh M. S. Khan, (Salman Khan dkk., 2015). Penelitian ini melakukan deteksi peristiwa jatuh berdasarkan bunyinya. Tujuan dari penelitian ini adalah mengelompokkan bunyi jatuh dan bukan jatuh. Pengelompokkan bunyi jatuh dan bukan jatuh menggunakan metode *one Class Support Vector Machine* (OCSVM) (Schölkopf dkk., 2001). Pada awalnya dilakukan esktraksi ciri menggunakan *Mel-Frequency Cepstral Coefficient* (MFCC) untuk kemudian dimodelkan dengan OCSVM, sehingga menghasilkan kelompok bunyi jatuh dan bukan jatuh. Metode ini tidak memerlukan data bunyi jatuh sebagai data pelatihannya, cukup dengan bunyi normal saja. Bunyi normal yang dimaksud adalah bunyi yang saling tumpang tindih antara bunyi jatuh dan bukan jatuh. Proses pengambilan data menggunakan *microphone* yang diletakan di suatu ruangan tertutup. Penggunaan ruangan tertutup mengakibatkan

bunyi gema ikut terekam sehingga tumpang tindih dengan bunyi aslinya. Bunyi gema menjadi *noise* pada data yang dikumpulkan. Bunyi *noise* dipisahkan atau direduksi dengan metode *spectral subtraction based binaural dereverberation method* (Khan dkk., 2013).

Penelitian lain yang juga menggunakan data bunyi lingkungan adalah penelitian tentang klasifikasi bunyi lingkungan, seperti pada penelitian Karol J. Piczak (2015). Penelitian ini melakukan klasifikasi jenis bunyi lingkungan, dengan melakukan esktraksi ciri pada masing-masing jenis bunyi untuk kemudian diuji melalui pengenalan atau klasifikasi secara manual (Piczak, 2015). Kemudian dilakukan klasifikasi dengan mesin menggunakan metode ekstraksi ciri *zero-crossing rate* dan MFCC, untuk pengenalannya menggunakan *random forest* dibandingkan dengan SVM atau kNN. Penelitian ini juga melakukan pengumpulan data bunyi lingkungan untuk kemudian dipublikasi dengan nama ESC-50 dan ESC-10, yang membedakan kedua kelompok tersebut adalah jenis dan jumlah datanya. Hasil klasifikasi manual, data ESC-10: 95,7% dan 81.3% untuk ESC-50. Hasil klasifikasi dengan mesin untuk data ESC-10 adalah 66.7% dengan kNN, akurasi 72.7% untuk *random forest*, dan 67.5% untuk SVM. Hasil untuk data ESC-50: 44.3% untuk *random forest*, 39.6% untuk SVM dan 32.2% untuk kNN (Piczak, 2015).

Data *non-speech* lainnya yang juga digunakan adalah data bunyi medis, seperti bunyi jantung dan paru seperti pada penelitian (Mondal dkk., 2017), (Chen dkk., 2017). Tujuan dari penelitiannya adalah memperkuat bunyi sinyal paru-paru yang sering kali tumpang tindih dengan bunyi jantung sehingga mengakibatkan *missing value* pada data bunyi paru-paru. Metode yang digunakan adalah melakukan prediksi pada *missing value* dengan algoritma prediksi berbasis *new Fast Fourier Transform* (FFT) yang kemudian dilakukan pengembalian kembali nilai sinyal bunyi paru-paru yang berdomain waktu menggunakan FFT *inverse* (Mondal dkk., 2017).

II.3 Sistem Deteksi Bunyi

Deteksi bunyi yang dimaksud pada subbab ini adalah suatu proses untuk mengenali jenis bunyi berdasarkan ciri bunyinya. Secara umum sistem deteksi jenis bunyi terdiri dari beberapa tahap, antara lain, pengumpulan data, *prepocessing*, ekstraksi ciri, penerapan teknik pemrosesan bunyi. Tahapan ini secara umum dilakukan pada setiap jenis bunyi dan tujuan dari pemrosesan bunyi, misalnya klasifikasi, identifikasi, maupun lokalisasi bunyi. Penjelasan dari langkah-langkah ini dijabarkan pada subbab-subbab berikutnya. Tahapan pengembangan sistem pengenalan bunyi dapat dilihat pada Gambar II.3.

Gambar II.3 Tahapan Pengembangan Sistem Pengenalan Bunyi

II.3.1 Pengumpulan data

Pada penelitian ini akan digunakan dua kelompok data bunyi yang telah dipublikasikan dan digunakan pada penelitian sejenis. Dua kelompok bunyi tersebut adalah *Environmental Sound Classification* (ESC) yang dipublikasikan pada tahun 2015, kelompok data bunyi kedua adalah TAU NIGENS 2020, data ini merupakan data yang digunakan pada *challenge* yang dibuat oleh kelompok peneliti yang berfokus pada pemrosesan bunyi, yaitu DCASE pada tahun 2020. Kedua kelompok bunyi ini akan digunakan pada pengembangan sistem pengenalan bunyi yang dikembangkan pada penelitian ini. Pembahasan detail dari kedua kelompok data akan dijabarkan pada bagian ini.

II.3.1.1 Kelompok data *Environmental Sound Classification* (ESC)

Data ESC merupakan kelompok data yang berisi potongan bunyi yang dibangun dari bunyi rekaman lainnya yang juga telah dipublikasi pada projek *Freesound*. Jenis label bunyi dari projek ini dipilih secara acak dengan tujuan mempertahankan keseimbangan dari jenis bunyi yang dominan. Hal ini dilakukan karena

keterbatasan jumlah dan keragaman jenis rekaman bunyinya. Penamaan label jenis bunyi dari *Freesound* berdasarkan jenis bunyi yang terdapat pada rekaman, misal bunyi kucing mengeong diberi label *cat* atau kucing. Proses anotasi label jenis bunyi dilakukan secara manual, artinya bunyi diputar dan didengarkan untuk kemudian diidentifikasi jenis bunyinya oleh seseorang. Hasil dari file audio atau bunyi ini memiliki format *sampling hertz* sebesar 44100 Hz, bersifat *single channel* dan dikompresi dengan *Ogg Vorbis* menjadi 192 kbit/s. Dataset yang telah diberi label kemudian disusun ke dalam lima lipatan validasi silang yang berukuran sama. Hasil kumpulan data ini dapat diakses secara non komersil dibawah naungan *Creative Commons* dengan lisensi oleh projek *Harvard Datarverse* (Piczak, 2015).

Kelompok data dari *Freesound* ini kemudian digunakan untuk membangun data ESC yang dibagi menjadi tiga kelompok data, kelompok data yang pertama adalah kelompok data utama yang disebut dengan ESC-50, kelompok data kedua adalah ESC-10 dan kelompok data terakhir adalah ESC-US. Kelompok data ESC-50 adalah kelompok data yang terdiri dari bunyi rekaman lingkungan sebanyak 2000 data audio yang secara seimbang terdiri dari 50 jenis kelas dengan masing-masing kelas memiliki 40 potongan rekaman berbeda. Data ini kemudian dikelompokkan menjadi lima kategori utama yaitu, suara hewan, bunyi lingkungan alami dan bunyi air, bunyi manusia yang *non-speech*, bunyi dari ruangan tempat tinggal yang tertutup dan kelompok terakhir adalah bunyi *noise* dari lingkungan tempat tinggal. Detail dari label kelompok jenis bunyi dapat dilihat pada Tabel II.1 (Piczak, 2014).

Tabel II.1 Kelompok jenis bunyi ESC-50.

Animals	***Natural soundcapes & water sounds***	***Human, non-speech sounds***	***Interior/domestic sounds***	***Exterior/urban noises***
Dog	*Rain*	*Crying baby*	*Door knock*	*Helicopter*
Rooster	*Sea waves*	*Sneezing*	*Mouse click*	*Chanisaw*
Pig	*Crackling fire*	*Clapping*	*Keyboard typing*	*Siren*
Cow	*Crickets*	*Breathing*	*Door, wood creaks*	*Car horn*
Frog	*Chirping birds*	*Coughing*	*Can opening*	*Engine*
Cat	*Water drops*	*Footsteps*	*Washing machine*	*Train*

Animals	***Natural soundcapes & water sounds***	***Human, non-speech sounds***	***Interior/domestic sounds***	***Exterior/urban noises***
Hen	*Wind*	*Laughing*	*Vacuum cleaner*	*Church bells*
Insects (flying)	*Pouring water*	*Brushing teeth*	*Clock alarm*	*Airplane*
Sheep	*Toilet flush*	*Snoring*	*Clock tick*	*Fireworks*
Crow	*Thunderstorm*	*Drinking, sipping*	*Glass breaking*	*Hand saw*

Kelompok data kedua yaitu ESC-10, kelompok data ini terdiri dari 10 kelas yang dipilih dari kelompok data ESC-50 yang merepresentasikan tiga kelompok bunyi umum yaitu kelompok bunyi pertama yaitu bunyi berpola secara temporal seperti suara bersin, anjing menggonggong, bunyi tik jam; kelompok kedua adalah kejadian bunyi dengan harmoni yang kuat contohnya suara tangisan bayi, suara ayam jantan berkokok; kelompok ketiga adalah kelompok jenis bunyi dengan *noise* yang lebih banyak atau sedikit, misalnya bunyi suara hujan, bunyi ombak di laut, bunyi benda terbakar, bunyi helikopter dan bunyi mesin gergaji. Kelompok bunyi ESC-10 ini muncul karena untuk mempermudah penerapan pada masalah yang lebih mudah. Misalnya ingin melakukan pengujian teknik klasifikasi bunyi dengan jumlah kelas yang terbatas yang sebenarnya mampu ditangani secara manual. Kebutuhan data ESC-10 ini lebih mudah digunakan karena memiliki perbedaan bunyi antar kelas yang jelas, tingkat ambiguitas yang rendah sehingga cocok diujikan pada berbagai jenis teknik pembelajaran mesin. Kelompok data ketiga adalah ESC-US, kelompok data ketiga ini muncul karena keterbatasan jumlah sampel dari masing-masing kelas, sehingga tidak cocok saat akan digunakan untuk permasalahan yang lebih kompleks, misalnya untuk mempelajari representasi dari data bunyi. Sehingga dilakukan penambahan data rekaman sebanyak 250.000 data yang diekstrak dari *Freesound*. Berbeda dengan ESC-50 dan ESC-10 yang dilakukan anotasi secara manual mengikut Freesound, data ESC-US ini memiliki metadata yang dilakukan dengan secara otomatis berdasarkan metadata anotasi manualnya.

II.3.1.2 Kelompok data TAU NIGENS 2020

Selain data ESC, kelompok data lainnnya juga digunakan pada penelitian ini yaitu TAU NIGENS 2020. Data ini digunakan pada *challenge* yang dibuat oleh kelompok studi DCASE. Data TAU NIGENS 2020 dibuat dari data publik lainnya yaitu data NIGENS. Data TAU-NIGENS 2020 merupakan data yang digunakan untuk menguji SELDnet. Data merupakan gabungan antara data teknik perekaman data bunyi TAU 2019 dengan data NIGENS. Teknik perekaman TAU 2019 digunakan pada TAU 2020, yang membedakan adalah jumlah kelas dan sampel yang digunakan. Proses perekaman dan format data keduanya sama yaitu FOA dan MIC, serta mencatat data *Azimuth* dan *Elevated* yang juga sama. Pembagian kelompok data juga sama yaitu terdiri dari enam kelompok yang dapat digunakan sebagai data latih dan uji. Sebelum membahas lebih lanjut tentang TAU 2020, akan dibahas terlebih dahulu data NIGENS, karena memang data TAU 2020 mensintesis datanya dari data NIGENS.

Nama NIGENS diambil dari *Neural Information processing group General Sounds.* Data NIGENS merupakan salah satu basis data dari bunyi dengan kualitas isolasi yang baik dengan ukuran data yang cukup besar yang dapat digunakan untuk melakukan simulasi kompleks dari bunyi serta untuk mengembangkan model deteksi bunyi yang bersifat *robust*. Isi dari data NIGENS berupa file stereo yang berformat wav berjumlah 1017 file dengan variasi durasi antara satu detik hingga lima menit, total durasi keseluruhan adalah 4 jam 45 menit. Sebagian besar data memiliki presisi 32-bit dan *sampling rate* sebesar 44100 Hz. Bunyi yang terekam adalah bunyi terisolasi tanpa bunyi latar dan gangguan lainnya. Kelas bunyi data NIGENS sebanyak 14 kelas, antara lain: alaram, tangisan bayi, tabrakan, gonggongan anjing, nyala mesin, nyala api, langkah kaki, ketukan pintu, suara laki-laki dan wanita berbicara, suara laki-laki dan wanita berteriak, dering telepon, dan kelas bunyi piano. Terdapat juga kelas bunyi umum yaitu bunyi-bunyi di luar 14 kelas lainnya, antara lain bunyi potongan kejadian maupun yang terus menerus bunyinya. Pada Lampiran A menunjukkan detail file dari masing-masing kelas, terdapat informasi jumlah file, total durasi, dan rata-rata durasi.

Seluruh RIR (*room impulse response*) multichannel terkestraksi dan sampel bunyi di-*sampel* ulang menjadi 24 kHz. Dari delapan bagian data set NIGENS yang tersedia, enam buah digunakan untuk membuat *development* dan dua data sisanya digunakan untuk kelompok data evaluasi. Satu atau dua ruangan ditentukan pada setiap bagian data dan 100 campuran dari bunyi spasial dihasilkan untuk setiap kombinasi dari sampel jenis bunyi dan ruangan. Panjang atau durasi dari setiap campuran data yang dihasilkan adalah satu menit. Permulaan dari bunyi pada setiap rekaman terdistibusi secara acak, tapi dibatasi oleh level *polyphonic* yang diijinkan, yaitu nilai minimalnya adalah satu dan nilai maksimalnya adalah dua.

Sebuah jenis bunyi dipilih secara acak baik statik maupun bergerak. Sumber bunyi yang statik ditentukan untuk DoA secara acak dari daftar referensi yang tersedia untuk ruangan yang spesifik. Bunyi bergerak ditentukan secara acak satu dari perekaman lintasan RIR dari suatu ruangan, oleh karenaya membatasi gerakannya bersamaan dengan jalurnya. Bagaimanapun juga arah gerakan dan tingkat gerakan dapat berbeda untuk setiap jenis bunyi. Arah gerakan diacak, sementara kecepatan gerakan dipilih secara acak dari tiga level yaitu lambat (~10 derajat/detik), menengah (~20 derajat/detik), dan cepat (~40 derajat/detik). Sebagai tambahan, karena setiap lintasan direkam pada ketinggian yang berbeda, bunyi bergerak mencapai akhir jalur memiliki kemungkinan untuk lompat ke elevasi yang lebih tinggi atau lebih rendah dan lanjut ke gerakannya pada jalur respektif dari tingginya.

Sumber bunyi statik dispasialkan oleh konvulasi dengan respektif RIR untuk DoA dan ditambahkan pada campuran datanya. Bunyi bergerak dispasialkan oleh konvulasi skema varian waktu yang dimunculkan antara STFT dari sampel bunyi dan STFT dari seluruh RIR sepanjang jalur geraknya. Operasi menyerupai sebuah skema konvulasi yang dipartisi, dengan RIR yang dikombinasikan dengan skema *cross fading* yang memberikan bobot lebih untuk frame dari RIR sebelumnya untuk ekor gema, dan bobot lebih untuk frame dari RIR terbarunya untuk arah langsung dan refleksi terkini. Oleh karena referensi DoA diekstrak pada sekitar interval 1 derajat disepanjang lintasan, kecepatan dari gerakan dikontrol menggunakan 10 (lambat), 20 (sedang), atau 40 (cepat) RIR berurutan per 1 detik keluarannya. Bunyi

yang sangat singkat yang lebih dari dua detik dikeluarkan dari menjadi dinamis dan ditentukan menjadi DoA statik. Setelah bunyi spasial dikonvulasi ditambahkan pada setiap gabungan multichannel dengan kesengajaan kondis *polyphonic*, gangguan (*noise*) lingkungan dari ruangan yang sama dicampur tambahkan. Bunyi gangguan lingkungan direkam yang dipisahkan menjadi segment satu menit dan ditambahkan pada campuran dengan beragam level signal to noise (SNR) dari antara 30 dB sampai 6 dB. Sebuah komponen segala arah didapat melalui sebuah kombinasi linier dari channel yang tanpa campuran *noise* dan rekaman *noise* gema.

Pada bagian data eksperimen yang digunakan, data terdiri dari beberapa jenis bunyi spasial yaitu jenis berdasarkan ruang akustiknya dan jenis data berdasarkan arah sumber bunyi dari alat rekam. Kondisi ruangan pengambilan data juga terdiri dari berbagai jenis bentuk ruangan, ukuran dan kondisi akustik serta peredam gema. Jenis sumber bunyi juga dibagi menjadi dua, yaitu sumber bunyi bergerak dan tidak bergerak atau statis. Setiap jenis bunyi terhubung dengan lintasan arah kedatangannya atau DOA (*Directions of Arrival*). Spesifikasi data yang berhasil dikumpulkan antara lain: *development set*: 600 menit/ 10 jam; *evaluation set*: 200 menit/ 3.33 jam; nilai *sampling rate* yang digunakan adalah 24000 Hz; kondisi *overlapping* dua atau lebih jenis bunyi; *noise* ditambahkan dengan kondisi: SNR: 30 dB-60dB. Data direkam di 15 ruangan tertutup (*indoor*) berbeda di Universitas Tampere, Finlandia, direkam juga 30 menit *noise* dengan lokasi dan kondisi perekaman yang sama. Sudut dari sumber bunyi juga dicatat sebagai parameter (DoA), menggunakan sudut *Azimuth* $ø \in [-180,180]$ dan sudut *Elevation* $\theta \in [-45,45]$. Data dibagi menjadi dua jenis format penyimpanan. Jenis format data yang pertama yaitu *first order ambisonics* (FOA), jenis kedua yaitu tethrahedral mic array (MIC) (Cao, dkk., 2019). Pada data FOA untuk data DoA menggunakan $H_m(ø, \theta, f)$ yang didapat dengan mengubah nilai *Azimuth* dan *Elevation* yang bersifat polar menjadi 3DCartesian. Sedangkan pada MIC akan diubah menggunakan Legendre polynomial untuk mencatat sudut cos antara *microphone* dengan DOA-nya. Kedua data tersebut disebut sebagai TAU 2020. Selain data, berikut adalah rincian dari perangkat pengembangan program dan eksperimen yang digunakan:

sebagai *compiler development tools* adalah *Visual studio code*, bahasa pemrograman yang digunakan adalah Python versi 3.7.3.

II.3.2 Ekstraksi fitur

Bunyi tentu memiliki ciri yang berbeda berdasarkan sumber bunyi dan waktu terjadinya bunyi. Ciri yang didapatkan harus dapat mewakili label atau kelompok bunyinya. Teknik yang dibangun juga harus menentukan ciri yang dapat membedakan suara yang satu dengan yang lain. Terdapat beberapa metode yang umum digunakan untuk mengekstraksi ciri bunyi yaitu *Discrete Fourier Transform* (DFT), *Fast Fourier Transform* (FFT) *dan Mel Frequency Cepstrum Coefficients* (MFCC), atau *Short Time Frequency Transform* (STFT). Pada penelitian yang dilakukanekstraksi ciri yang digunakan adalah STFT dan MFCC.

II.3.2.1 *Short Time Frequency Transform*

Implementasi dari Algoritma *Short Time Frequency Transform* (STFT) menggunakan *library librosa* berbasis Python *language*. Algoritma STFT yang diterapkan menggunakan pendekatan pemrosesan musik. Algoritma STFT menerapkan Algoritma *Fast Fourier Transform*, namun STFT melakukan proses *windowing*. *Windowing* dilakukan pada signal yang telah disegmenkan. Ukuran *window* dan segmen ditentukan di awal proses. Kemudian, proses FFT dilakukan pada hasil *windowing*-nya. Algoritma STFT yang digunakan dapat dilihat pada *pseudocode* berikut ini (Mueller, 2015):

Algoritma penerapan STFT

Input: Signal

Output: nilai STFT

Prosedur:

1. *Framing sinyal menjadi beberapa segment pendek*
2. *Setiap short segment dilakukan pengalian dengan window function, yang digunakan adalah Hann window*
3. *Hann window:* $h(t) = \frac{1}{2} + \frac{1}{2} cos \frac{2\pi n}{N}, \; n = -\frac{N}{2}, \dots, \frac{N}{2}$ *(II.1)*
4. *Dilakukan FFT untuk setiap segemen.*
5. *Dilakukan perhitungan STFT:*

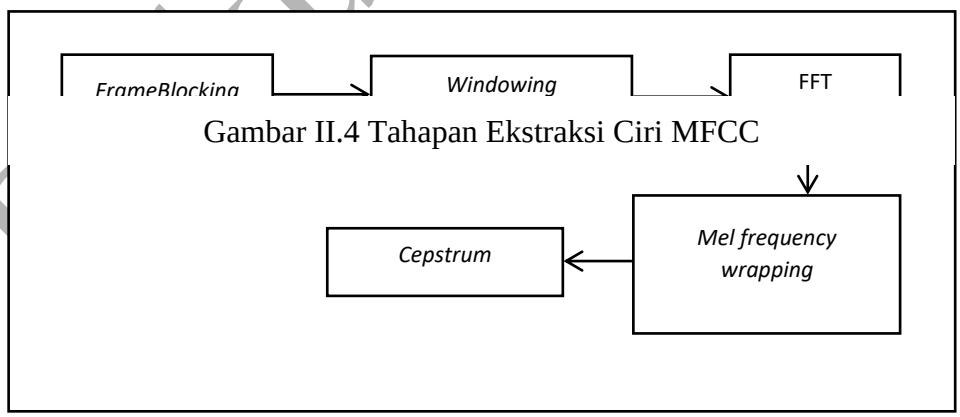
6. $X(n,k) = \sum_{m=0}^{N-1} x(m+nH)h(m)e^{-\frac{j2\pi km}{N}}$ *(II.2)*

di mana:

H: time step (hop size)
k: frekuensi $f(k) := \frac{kf_s}{N}$
n: waktu $t(n) := \frac{nH}{f_s}$

II.3.2.2 Metode *Mel Frequency Cepstrum Coefficients*

Metode *Mel Frequency Cepstrum Coefficients* (MFCC) merupakan metode ekstraksi ciri suara yang bertujuan untuk menangkap ciri suara berdasarkan sinyal-sinyal diskritnya (Ganchev, 2005). Sinyal diskrit yang berbasis waktu akan diubah menjadi berbasis frekuensi yang lebih mudah diteliti. Metode MFCC dikembangkan berdasarkan studi *psychophysical* yang menyebutkan bahwa suara manusia tidak bersifat linear, pada MFCC hal ini menjadi dasar untuk melakukan *filtering* dengan skala mel. Skala mel bersifat linear pada frekuensi suara bernilai lebih kecil dari 1000 Hz sedangkan di atas 1000 Hz bersifat logaritmik. Langkah selanjutnya adalah mengubah kembali spectrum log mel menjadi spektrum waktu menggunakan metode *Discrete Cosine Transform* (DCT) dan hasilnya disebut sebagai *Mel Frequency Cepstrum Coefficents*. Hasil dari MFCC merupakan ciri suara yang menjadi input pada proses selanjutnya. Tahapan MFCC dapat dilihat pada Gambar II.4.



Gambar II.4 Tahapan Ekstraksi Ciri MFCC

Tujuan dari *frame blocking* adalah melakukan segmentasi dari sinyal suara (Ganchev, 2005). Segementasi dilakukan karena kecepatan dari pengucapan kata tiap individu berbeda, dengan adanya segmentasi panjang data akan konsisten. Proses *frameblocking* akan dilakukan secara *overlapping* agar tingkat kontinuitas datanya tetap terjaga. Tahapan selanjutnya adalah *windowing*. Tujuan dari tahap

windwoing adalah untuk menghaluskan hasil pemotongan dari tahapan *frame blocking*. Proses *windowing* ini akan mencegah adanya perubahan data antar *frame* yang terlalu jauh kisaran nilainya (Ganchev, 2005). Metode yang umum digunakan adalah *Hamming window* (Ganchev, 2005). Data suara yang diambil berada pada domain waktu, hal ini mengakibatkan kesulitan pada proses perhitungan dan analisis. Proses MFCC pada tahap *Fast Fourier Transform* (FFT), mengubah domain waktu menjadi domain frekuensi sehingga akan lebih mudah untuk dianalisis (Ganchev, 2005).

Setelah data diubah menjadi domain frekuensi maka tahap selanjutnya adalah *mel frequency wrapping.* Tahap ini menyesuaikan frekuensi yang didapat dengan frekuensi suara manusia, sehingga akan menghasilkan ciri suara yang bersesuaian dengan suara manusia (Ganchev, 2005). Tahap terakhir dari ekstraksi ciri suara manusia adalah *cesptrum.* Data frekuensi diubah kembali menjadi data berdomain waktu. Hasil dari tahap *cepstrum* menjadi koefisien MFCC yang digunakan sebagai hasil ekstraksi ciri suara (Ganchev, 2005). Ukuran data akhir ekstraksi ciri suara adalah sebesar n x m, dimana n adalah jumlah dari data yang diinginkan (diinput pada tahap *cepstrum*), sedangkan m adalah panjang data yang didapat saat proses perubahan data analog menjadi data digital dan pada tahap *frame blocking*. Hasil dari MFCC ini menjadi input pada proses selanjutnya yaitu proses pelatihan dan pengenalan.

II.3.3 Teknik Pengenalan Bunyi

Pada pemrosesan bunyi terdapat beberapa tujuan antara lain klasifikasi bunyi yaitu melakukan pengelompokan bunyi berdasarkan ciri yang telah diekstraksi, identifikasi yaitu mengenali jenis bunyi yang terdengar dan lokalisasi yaitu mendeteksi arah datang bunyi yang bersumber dari sumber bunyi yang bergerak. Terdapat beberapa teknik yang digunakan pada penelitian ini yaitu *Support Vector Machine* (SVM) untuk klasifikasi bunyi, Teknik *Artificial Neural Network* (ANN) yang diterapkan pada *Sound Event Detection and Localization (SELDnet)* yang digunakan untuk identifkasi dan lokalisasi bunyi. Selain itu pada bagian ini juga

akan dibahas tentang teknik pemisahan bunyi yang menjadi bagian sistem yang dikembangkan.

II.3.3.1 *Support Vector Machine*

Metode *Support Vector Machine* (SVM) dapat diterapkan untuk melakukan klasifikasi (Gunn, 1998). Tujuan dari metode SVM adalah membentuk *optimal separating hyperplane*. Permasalahan klasifikasi dengan SVM yaitu memisahkan himpunan *training* dalam bentuk vektor menjadi dua kelas. Himpunan data latih awal dapat dinotasikan dengan (Gunn, 1998):

$$D = \{ (x^1, y^1), \dots, (x^1, y^1)\}, x \in R^n \;\; y \in \{-1,1\} \quad (\text{II.3})$$

dengan hyperplane (Gunn, 1998):

$$< w, x > + b = 0 \quad (\text{II.4})$$

Kondisi optimal jika: Terpisah tanpa *error*

Jarak maksimum antar vektor terdekat dengan hyperplane

Pada penelitian digunakan (Gunn, 1998):

$$\min \left| < w, x^i > + b \right| = 1 \quad (\text{II.5})$$

$w =$ bobot yang diperoleh dari perhitungan inverse jarak dari titik terdekat ke hyperplanenya.

II.3.3.2 Teknik k *Means Clustering*

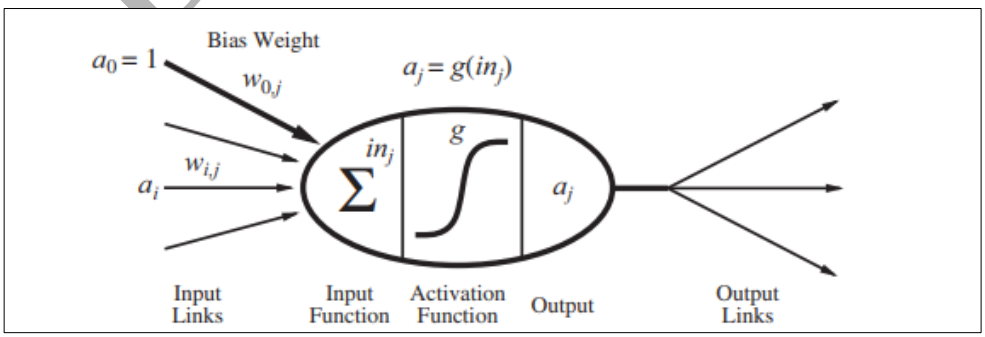
Teknik k Means *clustering* digunakan untuk memisahkan data matriks. Data matriks *activation* dan *template* masing-masing diklaster. Jumlah klaster adalah jumlah yang didapat pada tahap *sound recognition trigger*. Namun, pada pra eksperimen di SK I jumlah klaster ini masih diinputkan secara manual dan statis yaitu dua. Sedangkan pada sistem nanti rentang nilainya antara 2-3 sumber bunyi. Hasil dari klaster matriks *activation* dan *template* akan digabungkan kembali untuk membentuk data bunyi yang telah terpisah. Jumlah data bunyi yang terpisah adalah jumlah klasternya. Implementasi pengkodean yang digunakan pada sistem adalah *scikit-learn* berbasis *python language*. Adapun algoritma kMeans yang digunakan adalah sebagai berikut (Scikit-learn developer, n.d.) :

1. *Input: data yang ingin diklaster dalam matriks*

2. *Tentukan centroid data dari input, jumlah centroid berdasarkan jumlah klaster yang diinginkan*
3. *Optimasi anggota klaster berdasarkan jarak terdekat dari masing-masing centroid.*
4. *Output: hasil klaster berupa data-data yang telah terkelompok berdasarkan centroid-nya*

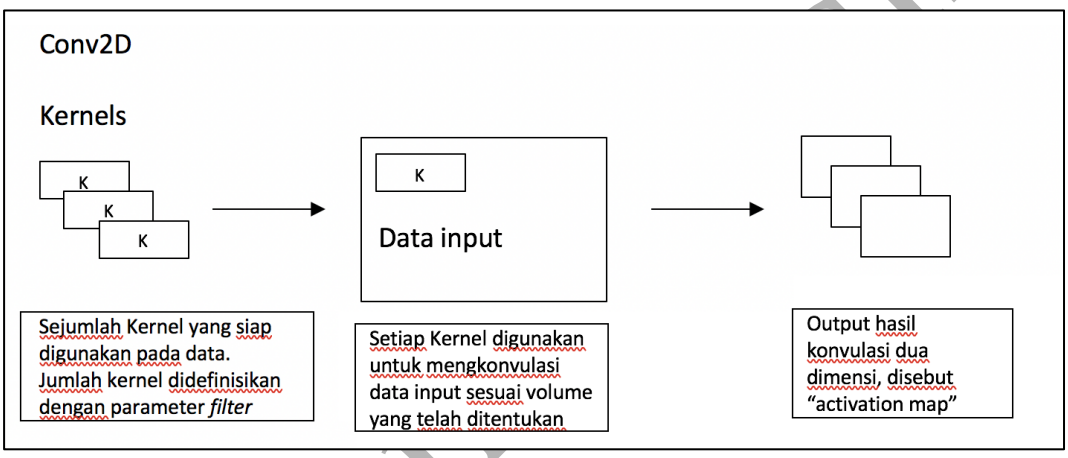
II.3.3.3 *Artificial Neural Network*

Jaringan syaraf manusia merupakan susunan neuron yang didalamnya terjadi aktivitas elektrokimiawi untuk mengatur aktivitas mental manusia (Russell dan Norvig, 2016). Teknik *Artificial Neural Network* (ANN) menggunakan pendekatan yang meniru cara kerja jaringan syaraf manusia tersebut. Misal, ingin menentukan suatu kondisi dikatakan kebakaran jika terdapat beberapa parameter atau syarat yang menentukan terjadi suatu kebakaran, misal terdapat asap, terdapat nyala api, terdapat bau gosong, dan lain sebagainya. Nilai dari parameter-parameter itu akan menentukan nilai terjadi atau tidak terjadi kebakaran. Secara matematika model untuk sebuah neuron pada jaringan syaraf tiruan dapat dilihat pada Gambar II.5. Model matematika tersebut memiliki aktivasi keluaran unit berupa $a_j = \left(\sum_{i=0}^{n} w_{i,j},\ a_i\right)$, dimana a_i adalah aktivasi output dari unit i dan $w_{i,j}$ adalah bobot pada hubungan dari unit i hingga ke unit aktivasi. Jadi, pada dasarnya sebuah jaringan syaraf tiruan adalah sebuah kumpulan unit yang saling terhubung satu dengan yang lain, properti dari jaringan ditentukan oleh topologi dan properti dari "neuron" (Russell dan Norvig, 2016).



Gambar II.5 Model matematika sederhana dari sebuah neuron di jaringan syaraf tiruan (Russell dan Norvig, 2016).

Teknik CNN merupakan salah satu pendekatan pada ANN yang diterapkan pada beberapa penelitian terkait bunyi lingkungan, antara lain. Pada dasarnya CNN juga terdiri dari neoron yang memiliki bobot, bias, dan fungsi aktivasi Terdapat beberaja jenis CNN yang telah dikembangkan, salah satunya adalah CNN *Convolutional Two Dimensional* (Conv2D). Teknik Conv2D merupakan teknik yang menggunakan kernel konvulasi. Kernel ini akan mengkonvulasikan data input. Disebut konvulasi dua dimensi, karena memang kernel yang digunakan berukuran dua dimensi, yaitu lebar dan tinggi. Kernel yang digunakan bisa lebih dari satu, jumlah kernel yang digunakan adalah *filter* pada penelitian SELDnet. Keluaran dari konvulasi juga berupa dua dimensi yang disebut sebagai *activation map*.



Gambar II.6 Tahapan Conv2D

II.3.3.4 SELDnet

Deteksi dan lokalisasi peristiwa bunyi merupakan dua tujuan penelitian dari identifikasi aktivitas temporal untuk jenis bunyi. Bunyi yang aktif akan diestimasi lokasi spasialnya, inilah yang disebut dengan deteksi dan lokalisasi bunyi. Jadi, pada bagian ini akan dijelaskan dua tujuan penelitian yaitu deteksi jenis bunyi dan lokalisasi sumber bunyi. Penelitian deteksi jenis bunyi melakukan pengenalan bunyi berdasarkan jenis bunyinya disebut dengan *Sound Event Detection* (SED). Penelitian SED telah banyak dilakukan dengan metode *supervised classification* yang melakukan pengenalan per *frame* serta dengan kondisi bunyi yang saling tumpang tindih. Pada beberapa penelitian lain bukan hanya menggunakan berbagai teknik pengenalan tapi juga menggunakan format data berdasarkan alat rekam yang berbeda untuk meningkatkan hasil akurasi deteksi. Format data yang umum

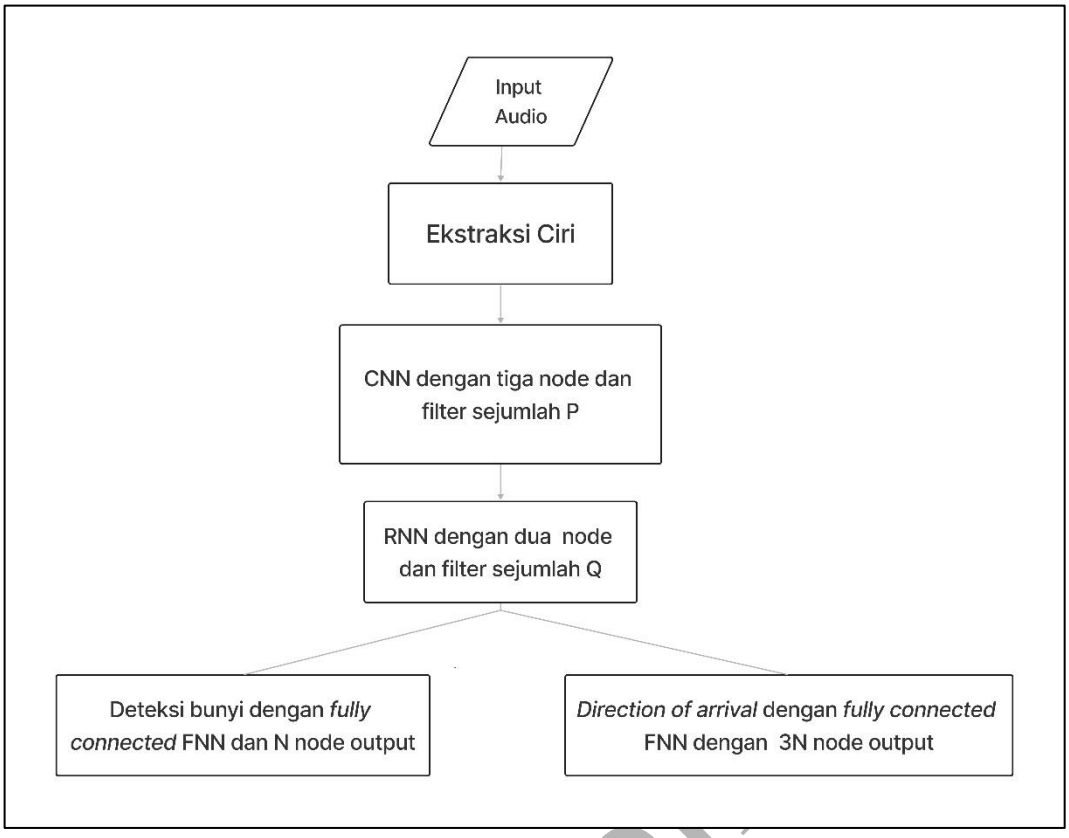
digunakan adalah *michrophone* tunggal, *binaural* dan *first-order Ambisonics* (FOA) (Adavanne, dkk., 2018).

Setelah deteksi jenis bunyi dilakukan, maka lokalisasi bunyi dilakukan pada sistem yang dibangun. Pada beberapa penelitian melakukan deteksi secara simultan berdasarkan arah kedatangan sumber bunyi atau disebut sebagai *direction-of-arrival* (DOA). DOA inilah yang dijadikan acuan untuk melakukan lokalisasi bunyi- (Adavanne, dkk., 2017; Chakrabarty dan Habets, 2017; He, dkk., 2017; Hirvonen, 2015). Secara umum terdapat dua teknik yang digunakan untuk DOA, yaitu *parametic* dan *Deep Neural Network* (DNN). Teknik *parametic* yang dikembangkan antara lain *time-difference-of-arrival* (TDOA) (Huang, dkk., 2001), *steered-response-power* (SRP) (Dibiase, 2000), *multiple signal classification* (MUSIC) (Schmidt, 1986), dan *estimation of signal parameters via rotational invariance technique* (ESPRIT) (Roy dan Kailath, 1989). Namun, teknik *parametic* memiliki keterbatasan dalam menangani data yang memiliki *noise* berupa gema dan nilai *signal-to-noise* (SNR) yang rendah. Sedangkan, teknik DNN dapat lebih unggul menangani data dengan gema dibanding teknik *parametic* (Adavanne, dkk., 2017; Chakrabarty dan Habets, 2017; He, dkk., 2017; Hirvonen, 2015). Selain itu, penelitian DOA dengan data yang saling tumpang tindih dengan mengestimasi jumlah sumber bunyi yang aktif juga telah dilakukan dengan teknik klasifikasi regresif dan DNN (Ferguson, dkk., 2018; Vesperini, dkk., 2016). Selain berdasarkan teknik dan alat rekam, performa deteksi dan lokalisasi bunyi dioptimasi menggunakan berbagai jenis geografis pada alat rekam, yaitu *full azimuth* dan *linear arrays*. Pada *linear arrays* sudut yang digunakan adalah 180 derajat. Terdapat juga penggabungan antara *full azimuth* dan *linear arrays* menggunakan FOA (*first order ambisonics)* atau sinyal *Ambisonics* (Teutsch, 2007).

Kedua tujuan penelitian, deteksi dan lokalisasi bunyi, yang masing-masing telah banyak diteliti dan dikembangkan. Penggabungan kedua tujuan penelitian tersebut dimaksudkan untuk berbagai tujuan, terutama pada sistem *monitoring* yang akan dibangun. Terdapat beberapa metode penggabungan teknik deteksi dan lokalisasi

bunyi yang memiliki tingkat akurasi tinggi, antara lain HIRnet (Hirvonen, 2015) dan SELDnet. HIRnet lebih unggul pada data *speech* dan musik, sedangkan SELDnet unggul pada data bunyi lingkungan, sehingga SELDnet yang akan digunakan pada sistem *monitoring*.

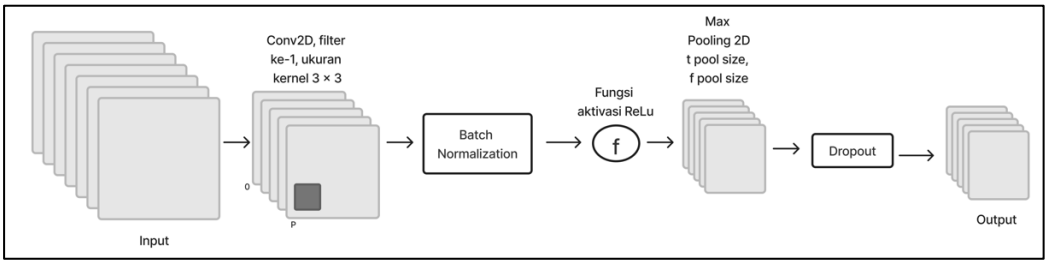
Pada SELDnet menerapkan gabungan teknik klasifikasi antara *Convolutional Neural Networks* (CNN), *Recurrent Neural Netwroks* (RNN) dan *fully connected* (FC) dengan *layers* berurut. CNN digunakan sebagai tahap awal setelah ekstraksi ciri dengan multiple layer 2D CNN dengan tiga layer (Adavanne, dkk., 2018; Adavanne, Pertilä, dkk., 2017; Çak, dkk., 2017; Lim, dkk., 2017; Scenes, 2017). Kemudian tahap selanjutnya adalah RNN yang memang tangguh untuk data sekuensial (Adavanne, Pertilä, dkk., 2017). Detail dari kerangka kerja dapat dilihat pada Gambar II.7. Pada Gambar II.7 atas merupakan bagan SELDnet (*Sound Event Localization and Detection*) yang digunakan untuk melakukan deteksi bunyi dan lokalisasi jenis bunyi yang dibangun oleh Sharath, A., dkk (2018). SELDnet dirancang untuk melakukan deteksi dan lokalisasi peristiwa bunyi yang terjadi secara tumpang tindih dan simultan. Keunggulan yang dinyatakan pada penelitian ini adalah dapat menangani kondisi tumpang tindih yang lebih dari dua jenis bunyi, tidak seperti pada penelitian yang lain yang masih terbatas hanya pada dua jenis bunyi tumpang tindih, padahal pada kondisi nyata bunyi tumpang tindih bisa lebih dari dua jenis. Keunggulan lainnya adalah SELDnet dapat menangani kondisi bunyi dengan kondisi gema, gema sering kali menjadi kelemahan dari sistem pengenalan bunyi dan dianggap sebagai *noise* yang pada akhirnya akan dihilangkan atau direduksi untuk meningkatkan hasil deteksi.



Gambar II.7 Kerangka kerja SELDnet (Adavanne, dkk., 2019)

Pada Gambar II.7 digambarkan bahwa sistem diawali dengan menerima input audio. Kemudian diekstrak nilai spectogramnya untuk setiap *channel C* dari *multichannel audio* menggunakan *discrete Fourier Transform* (DFT) dengan tipe *window Hamming* yang panjangnya *M* dan *overlap*-nya sebesar 50 persen. Keluaran berupa nilai *fase spectrogram* dan maginitud akan digunakan sebagai fitur. Namun, hanya nilai frekuensi positif tanpa nilai nol dari *M/2* yang digunakan sebagai fitur. Pada Gambar II.7 bagian *Feature extractor* menghasilkan keluaran dengan ukuran dimensi *TxM/2x2C* yang menunjukkan bahwa fitur berupa sekuen *T frames* dengan dimensi *2C* yang merupakan komponen *C magnitude* dan *C phase*. Keluaran dari proses *feature extraction* digunakan sebagai input pada tahap *neural network*. Namun, tidak langsung keseluruhan fitur pada spectrogram yang digunakan, tapi fitur lokal variant yang digunakan sebagai pelatihan menggunakan *multiple layers* 2D CNN, seperti yang disimulasikan pada Gambar II.6. Setiap layer CNN memiliki filter *P* dengan dimensi *3 x 3 x 2C* dengan fungsi aktivasi yang digunakan adalah ReLU (*Rectified Linear Unit*). Jumlah layer CNN yang

digunakan sebanyak tiga buah, detail per satu layer CNN dapat dilihat pada Gambar II.8.



Gambar II.8 Topologi per satu layer CNN

Setelahnya, keluaran aktivasi akan dinormalisasikan menggunakan *batch normalization* dan pengurangan dimensi menggunakan *max-pooling* (*M,Pi*) sepanjang sumbu axis frekuensi. Dimensi keluaran setelah layer akhir CNN dengan filter *P* adalah *T x 2 x P*. Aktivasi *output* dari CNN selanjutnya diubah ukurannya menjadi urutan *frame* berukuran *T* dengan panjang fiturnya 2P, untuk kemudian menjadi input pada layer RNN. Pada RNN menggunakan layer *Gated Recurrent Unist* (GRU) dan aktivasi *tanh*, serta dilakukan *bidirectional*. RNN sendiri merupakan bagian dari *neural network* yang tangguh untuk data sekuen atau *time series data*, sehingga cocok digunakan pada SELDnet ini. GRU merupakan salah satu jenis RNN yang melakukan proses pelatihan yang lebih cepat dibandingkan jenis RNN lainnya misal *Long Short-Term Memory* (LSTM) (Chung, dkk., 2014). Setelah RNN didapatkan akan dibagi menjadi dua bagian yaitu untuk SED dan DOA, keduanya dilakukan *Fully Connected* (FC) dengan jumlah filter node sebanyak *R*. Pada SED menggunakan fungsi aktivasi sigmoid dengan jumlah filter node *N*, sedangkan pada DOA dengan tiga node (3N) yaitu *x, y, z* adalah data arah bunyinya. Keluaran dari SED akan berupa hasil klasifikasi dengan *multi-lable* berupa nilai kontinu [0,1] dan keluaran dari DOA berupa estimasi regresi dengan *multi-output* berupa nilai kontinu [-1,1].

Proses pelatihan dengan menentukan target awal DOA dan SED. Pada DOA target awal berupa x, y, dan z yang digunakan saat bunyi aktif, ketiganya akan bernilai nol saat bunyi tidak aktif. Nilai target SED sendiri akan bernilai satu jika bunyi sedang terjadi atau aktif, dan akan bernilai nol saat bunyi tidak aktif. Keluaran SED

berupa label kelas bunyi lingkungan, jika bunyi aktif maka bar pada garis waktu akan muncul sesuai dengan label kelasnya. Terjadi bunyi tumpang tindih antara bunyi suara manusia dengan bunyi gonggongan anjing pada waktu *frame t*. Pada keluaran DOA, labelnya berupa nilai posisi *x, y, z* yang nilainya [-1.1], hasilnya pada masing-masing kelas akan memiliki nilai posisinya berdasarkan waktu tertentu. Pada DOA juga dapat dilihat bunyi tumpang tindih yang aktif pada setiap posisinya dengan waktu *frame t.* Hasil menunjukkan kelas bunyi dengan nilai SED yang dimana jika nilainya diatas 0.5, maka bunyi akan dianggap aktif. Pada bagian *DOA Estimatesi* menunjukkan keluaran dari estimasi ruang 3D berkontinu yang diperoleh dari regresi dengan keluaran lebih dari satu. Nilai estimasi berupa koordinat *x, y, z* 3D *Cartesian* dari satuan sekeliling mikrofon.

II.4 Pemisahan bunyi tumpang tindih

Terdapat dua jenis teknik pemisahan yang akan digunakan pada penelitian ini. Keduanya diujikan pada eksperimen yang berbeda karena kedua teknik ini memiliki pendekatan yang berbeda. Kedua teknik yang diujikan adalah *Nonnegative Matrix Factorzation* (NMF) sebagai jenis *blind separation*, dan jenis kedua adalah *Time Frequency* (T-F) *masking* sebagai jenis *nonblind separation.*

II.4.1 *Non-negative matrix factorization*

Setelah esktraksi ciri bunyi didapat, selanjutnya adalah menerapkan teknik Metode *Nonnegative Matrix Factorzation* (NMF) untuk pemisahan bunyi tumpang tindih. Teknik NMF banyak digunakan pada pengolahan citra digital dan bunyi dalam hal kompresi dan perolehan ciri. Beberapa contoh penggunaan NMF pada penelitian citra digital adalah NMF sebagai representasi citra digital (Liu dkk., 2012), penggunaan NMF dalam verifikasi citra wajah (Zafeiriou dkk., 2006), NMF juga digunakan untuk pemisahan data citra yang saling tumpang tindih (Rajabi and Ghassemian, 2014) (Niranjani and Vani, 2017). Inti dari metode NMF ini adalah melakukan proses faktorisasi terhadap suatu matriks. Pada pengolahan citra dan bunyi data bit keduanya dijadikan atau dipandang sebagai satu matriks dua dimensi. Matriks inilah yang akan difaktorkan. Hasil dari faktor menjadi bagian dari hasil kompresi dan perolehan ciri. Metode NMF juga banyak digunakan pada

pengolahan suara dan bunyi dengan tujuan untuk melakukan pemisahan data bunyi saling tumpang tindih.

Salah satu penelitian yang menggunakan NMF adalah (Tripathi and Baruah, 2017). Penelitian ini melakukan deteksi *sound event* pada kondisi nyata menggunakan supervised *Nonnegative Matrix Factorzation* (NMF). Melakukan separasi dan klasifikasi pada single channel bunyi lingkungan. Metode yang digunakan pada ekstraksi ciri adalah kombinasi *Common Fate Transformation* dan *Cauchy Non-negative Matrix Factorization* sedangkan tahapan klasifikasi menggunakan *Fuzzy rule-based classifier*. Pembandingnya dengan metode SVM dengan data *real time* yang sama. Namun muncul tantangan untuk bunyi lingkungan sinyalnya tidak terstruktur serta terdiri dari berbagai macam sumber bunyi dan deteksi bunyi suara manusia atau *speech* lebih mudah dilakukan karena sinyal dan polanya yang terstruktur. Dengan kata lain, pemisahan bunyi untuk suara manusia dengan non suara manusia telah berhasil dilakukan, namun pemisahan data bunyi yang tumpang tindih non suara manusia atau bunyi lingkungan saja masih dilakukan dengan proses *tagging* secara manual.

Metode NMF juga digunakan pada penelitian di bidang medis untuk bunyi kardiak dan pernapasan (Shah dkk., 2015). Pada bidang medis, bunyi kardiak dan bunyi pernapasan terekam saling tumpang tindih, sehingga memerlukan ketelitian dalam melakukan diagnosa berdasarkan bunyi keduanya (Shah dkk., 2015). Pemisahan bunyi jantung dan pernapasan telah menjadi tujuan penelitian yang umum dilakukan. Pada penelitian oleh G. Shah, dkk. (2015) menggunakan metode NMF yang telah dimodifikasi sebagai metode untuk memisahkan bunyi kardiak dengan bunyi pernapasan. Hasil penelitiannya menunjukkan metode NMF yang telah dimodifikasi ini menghasilkan akurasi pemisahan yang tinggi bahkan pada kondisi lingkungan dengan *noise*.

Untuk dapat memahami secara sederhana NMF dapat dimulai dengan memahami pemfaktoran suatu bilangan. Jika suatu bilangan bulat positif ingin difaktorkan,

dapat dilakukan pemfaktoran secara sederhana, misal suatu bilangan bulat positif 18 ingin difaktorkan, maka hasilnya:

$$18 = 2 \cdot 3^2$$

atau, bilangan lain seperti 20, maka:

$$20 = 2^2 \cdot 5$$

serta contoh lainnya. Namun, perlu dicari tahu jika suatu matriks yang ingin difaktorkan.

Misal terdapat vektor berdimensi n:

$$x_I = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{pmatrix}; x_{II} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \dots, x_m = \begin{pmatrix} z_1 \\ z_2 \\ \dots \\ z_n \end{pmatrix} \quad (\text{II.6})$$

Kemudian akan diubah ke dalam sebuah matriks V:

$$V = \begin{bmatrix} x_1 & y_1 & & z_1 \\ x_2 & y_2 & \dots & z_2 \\ \dots & \dots & & \dots \\ x_n & y_n & & z_n \end{bmatrix} \quad (\text{II.7})$$

Maka ukuran dari matrik V adalah n x m, di mana n adalah dimensinya dan m adalah banyaknya sampel pada data. Selanjutnya adalah mencari tahu cara memfaktorkan matriks V (Lee dkk., 2001):

$$V \approx W \cdot H \quad (\text{II.8})$$

Ukuran dari matriks W adalah n x r, sedangkan matriks H adalah r x m, dimana ukuran kedua matriks tersebut lebih kecil dari pada matriks V.

Hasil pemfaktoran dilakukan secara iterasi untuk mendapatkan hasil pemfaktoran yang optimal. Pengulangan dilakukan dengan mengukur hasil pemfaktoran, pengukuran dilakukan dengan cost function, salah satu cara sederhananya adalah dengan menggunakan pengukuran jarak *Euclidean* pada rumus (II.9) (Lee dkk., 2001):

$$|A - B|^2 = \sum_{ij} \left(A_{ij} - B_{ij}\right)^2 \quad (\text{II.9})$$

Pencarian estimasi berawal pada batas bawa nol, dan akan berhenti jika A=B. Salah satu pengukuran yang juga digunakan dengan rumus (II.10) (Lee dkk., 2001):

$$D(A\|B) = \sum_{ij}\left(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}\right) \quad \text{(II.10)}$$

Sama seperti rumus (II.9) batas bawah (II.10) juga nol dan berhenti saat A=B. Nilai (II.10) tidak dapat disebut sebagai jarak karena nilai A dan B tidak memiliki arah tapi disebut sebagai divergen. Berdasarkan kedua rumus (II.9) dan (II.10), maka terdapat dua permasalahan atau problem yang ditetapkan untuk mencari NMF pada (II.8) yaitu (Lee dkk., 2001):

Problem 1. $Minimize\ \|V - WH\|^2$ dengan *W, H>0*

Problem 2. $Minimize\ D(V\|WH)$ dengan *W, H>0*

Pemecahan kedua problem di atas tidak dapat dipecahkan dengan pencarian nilai global minima, karena hasilnya minimalnya bersifat konveks hanya pada salah satu W atau H, untuk itu yang dicari adalah nilai local minimal dengan beberapa teknik. Pada penelitian yang diacu menggunakan teknik *Multiplicative update rules* (Lee dkk., 2001):

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \quad \text{(II.11)}$$

$$W_{ia} \leftarrow W_{ia}\frac{(V H^T)_{ia}}{(WHH^T)_{ia}} \quad \text{(II.12)}$$

Rumus (II.11) dan (II.12) digunakan jika menggunakan *Euclidean distance* $\|V - WH\|^2$ (Lee dkk., 2001).

$$H_{a\mu} \leftarrow H_{a\mu}\frac{\sum_i W_{ia}\, V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \quad \text{(II.13)}$$

$$W_{ia} \leftarrow W_{ia}\frac{\sum_\mu H_{a\mu}\, V_{i\mu}/(WH)_{i\mu}}{\sum_v H_{av}} \quad \text{(II.14)}$$

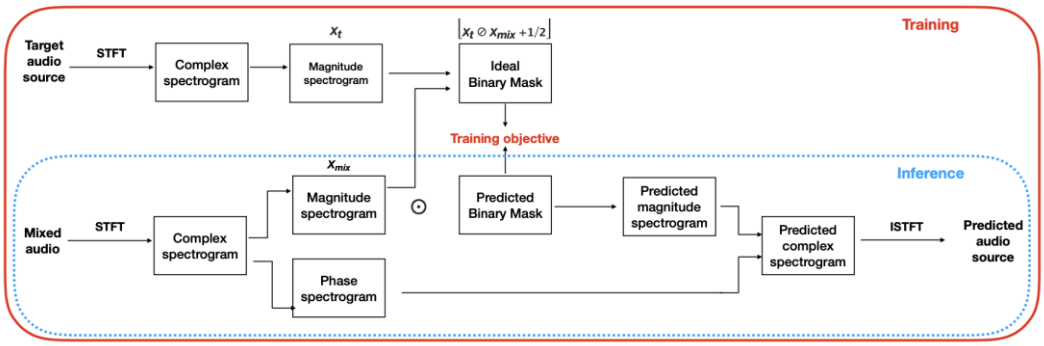
Rumus (II.13) dan (11.14) digunakan jika menggunakan divergen $D(V\|WH)$ (Lee dkk., 2001).

II.4.2 *Time-Frequency Masking*

Pemisahan bunyi yang dilakukan dengan *Time-Frequency Masking* (T-F) *masking* adalah teknik pemisahan yang berlawanan dengan teknik *blind separation* seperti NMF. Hal ini dikarenakan T-F *masking* menggunakan data latih untuk melakukan

pemisahan bunyi tumpang tindih. Hasil akurasi dengan teknik *masking* masih lebih unggul dibanding teknik pemisahan dengan *blind separation* (Luo dan Mesgarani, 2019). Teknik pemisahan T-F *masking* diawali dengan mencari nilai STFT dari data audio target dan tumpang tindih. Kemudian dari nilai STFT ini akan dihasilkan nilai magnitudnya. Data *masking* yang akan digunakan untuk pemisahan akan dihitung dari data spectrogram ini. Sehingga hasil pemisahan akan berupa nilai STFT yang kemudian akan dikonversi menggunakan inverse STFT untuk kemudian dijadikan data audio kembali.

Pada penelitian ini teknik T-F *masking* yang digunakan mengacu pada penelitian yang melakukan pemisahan bunyi antara bunyi musik latar dengan suara vokalnya. Proses pemisahan dilakukan dengan melatih data bunyi alat musik yang akan dipisahkan menggunakan T-F *masking*. Salah satu penerapan pemisahan bunyi tumpang tindih adalah *Northwestern University Source Separation Library* atau NUSSL. *Library* ini dibangun menggunakan bahasa pemrograman *Python*. Proses pemisahan diawali dengan pemrosesan dasar sinyal, seperti membaca data audio, *padding, adding*, dan memotong data audio, transformasi dan membuat data audio. Keseluruhan pemrosesan data audio di-*bundle* pada objek *AudioSignal*. Data audio *input* dan *output* berada pada dua dimensi data yang direpresentasikan menggunakan *array of pulse-code* (PCM). Kedua dimensi merepresentasikan waktu dan *channel*. Pada penelitian ini jumlah *channel* yang digunakan sebanyak empat buah mengacu pada data TAU NIGENS 2020. Proses dari pemisahan menggunakan T-F *masking* dapat dilihat pada Gambar II.9.



Gambar II.9 Tahapan Pemisahan T-F Masking (Yang dan Lerch, 2020)

II.5 Matrik evaluasi

Hasil pengujian eksperimen akan diukur menggunakan teknik pengukuran yang telah dilakukan pada penelitian acuan. Hasil teknik pemisahan dengan NMF akan diuji menggunakan klasifikasi SVM yang diukur nilai C nya. Semakin besar nilai C nya menandakan klasifikasi semakin baik. Sedangkan pada pemisahan T-F dan SELDnet akan menggunakan matriks evaluasi SELDnet yang akan dibahas pada subbab selanjutnya.

II.5.1 Matriks evaluasi SELDnet

Hasil eksperimen diukur menggunakan matriks skor yang dihitung berdasarkan nilai *F_score* dan *Error rate*-nya. Matriks skor dibagi menjadi dua bagian, bagian pertama yaitu *class-aware localization* dan *location-aware detection* serta bagian kedua yaitu *localization-only* dan *detection-only*. Pada matriks pertama *location-aware detection* untuk deteksi jenis bunyi hanya dengan hasil lokalisasi bunyi yang tepat, yang digunakan untuk perhitungan skor deteksi kelas bunyinya, sedangkan pada bagian kedua *detection-only*, tanpa melihat informasi lokasinya, tapi langsung menghitung skor akurasi deteksi bunyi. Begitu juga untuk lokalisasi bunyi, pada perhitungan matriks pertama, informasi ketepatan deteksi kelas bunyi disertakan, sedangkan pada bagian kedua tidak. Pada perhitungan skor untuk SED baik untuk yang *localization-aware detection* maupun yang *detection-only* menggunakan *error rate* dan *F_score*. Nilai *F_score* didapat dengan memperhatikan nilai $TP_{(k)}$ atau *true positive*, yang artinya hasil prediksi sama dengan data rujukannya, juga memperhatikan nilai $FP_{(k)}$ atau *false positive* dimana hasil prediksi menunjukkan kelas bunyi sedang aktif tapi pada data rujukkan tidak aktif, selain itu juga menggunakan nilai $FN_{(k)}$ yang merupakan *false negative*, misal hasil prediksinya menyatakan bahwa kelas bunyi tidak aktif, tapi data rujukannya adalah aktif. Persamaan (1) menunjukkan perhitungan *F_score*. Nilai *F_score* yang diharapkan 100 untuk performa terbaiknya (Adavanne, dkk., 2019).

$$F_{score} = \frac{2 \cdot \sum_{k=1}^{K} TP_{(k)}}{2 \cdot \sum_{k=1}^{K} TP_{(k)} + \sum_{k=1}^{K} FP_{(k)} + \sum_{k=1}^{K} FN_{(k)}} \quad \text{(II. 1)}$$

Perhitungan *Error Rate* untuk *localiation aware detection* diperoleh dari total jumlah *substitution* $S_{(k)}$, *deletion* $D_{(k)}$ dan *insertion* $I_{(k)}$ dibagi dengan total

keseluruhan kelas bunyi yang aktif dari data referensinya $N_{(k)}$, sehingga didapat persamaan (2).

$$Error\ Rate = \frac{\sum_{k=1}^{K} S_{(k)} + \sum_{k=1}^{K} D_{(k)} + \sum_{k=1}^{K} I_{(k)}}{\sum_{k=1}^{K} N_{(k)}} \quad \text{(II.2)}$$

Keterangan:

$$S_{(k)} = min(FN_{(k)}, FP_{(k)})$$

$$D_{(k)} = \max(0, FN_{(k)} - FP_{(k)})$$

$$I_{(k)} = \max(0, FP_{(k)} - FN_{(k)})$$

Nilai $FN_{(k)}$merupakan *false negative* misal, hasil prediksinya menyatakan bahwa kelas bunyi tidak aktif, tapi data rujukannya adalah aktif. Sedangkan nilai $FP_{(k)}$ adalah *false positive* artinya hasil prediksi menyatakan kelas bunyi aktif tapi pada data rujukan menunjukkan tidak aktif. Kedua nilai ini yang digunakan untuk menentukan nilai *error rate*. Nilai *error rate* 0 (nol) yang menunjukkan performa terbaik.

Pada perhitungan performa untuk DOA selain menggunakan *F_score*, perhitungan *DOA_error* juga digunakan sebagai pengukuran. Perhitungan *DOA_error* didapat dengan membandingkan data lokasi antara data hasil prediksi (x_E, y_E, z_E) dengan data acuan atau referensi (x_G, y_G, z_G) menggunakan rumus (3) dan (4).

$$\sigma = 2 \cdot \arcsin\left(\frac{\sqrt{\Delta x^2 + \Delta y^2 + \Delta z^2}}{2}\right) \cdot \frac{180}{\pi} \quad \text{(II. 3)}$$

di mana: $\Delta x = x_G - x_E$; $\Delta y = y_G - y_{E;}\, \Delta z = z_G - z_E$

$$DOA_{error} = \frac{1}{D} \cdot \sum_{d=1}^{D} \sigma\left(\left(x_G^d, y_G^d, z_G^d\right), \left(x_E^d, y_E^d, z_E^d\right)\right) \quad \text{(II. 4)}$$

di mana *D* adalah total jumlah estimasi DOA dari keseluruhan data set; $\left(\left(x_G^d, y_G^d, z_G^d\right), \left(x_E^d, y_E^d, z_E^d\right)\right)$ adalah sudut antara estimasi dan acuan DOA pada data ke-*d*.

Frame recall adalah perhitungan performa yang digunakan jika jumlah *frame* waktu antara estimasi dan acuan DOA tidak sama. Perhitungan *Frame recall* menggunakan rumus (II.6).

$$Frame\_recall = TP/(TP + FN) \quad \text{(II.5)}$$

Pada tahap pelatihan SELDnet untuk menentukan henti latih awal menggunakan perhitungan gabungan antara SED *score* dan DOA *score*.

$$SELD\_score = (SED\ score + DOA\ score)/2 \quad \text{(II. 6)}$$

di mana

$$SED\ score = (Error\ Rate + (1 - F_{score}))/2$$

$$DOA\ score = (DOA\_error/180 + (1 - Frame\_recall))/2$$

Perhitungan *SELD_score* menunjukkan performa SELDnet secara keseluruhan. Nilai optimal yang diharapkan pada *SELD_score* adalah nol, semakin mendekati nilai nol artinya performanya semakin baik.

II.6 Augmentasi Data Bunyi

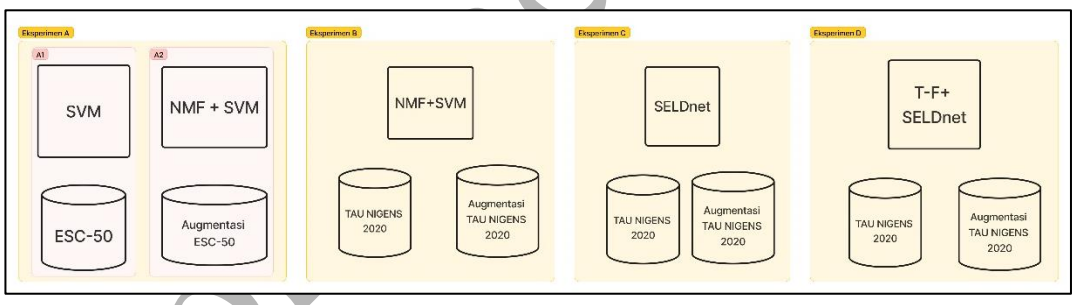
Proses augmentasi merupakan salah satu cara untuk memproduksi kumpulan data yang digunakan untuk keperluan eksperimen. Pada penelitian tentang bunyi proses augmentasi data menjadi lazim dilakukan guna menghasilkan data dengan lebih banyak variasi dan kondisi data bunyi yang dinginkan sesuai tujuan dan keperluan penelitian. Data bunyi sendiri secara umum dapat diperoleh dari proses perekaman langsung dari sumber bunyinya, seperti pada data NIGENS dan ESC-50, detail dari data ini telah dibahas pada subbab sebelumnya. Bunyi direkam dengan berbagai durasi dan kelompok sumber bunyi yang telah ditentukan. Hasil dari rekaman asli digunakan untuk membuat atau mengaugmentasi data bunyi baru dengan kondisi yang berbeda. Kondisi yang dimaksud adalah bunyi yang saling tumpang tindih atau bisa juga bunyi dengan kondisi spasial, seperti pada data TAU NIGENS 2020. Data TAU NIGENS 2020 diaugmentasi dengan cara memutar data bunyi NIGENS pada suatu ruangan dengan kondisi gema tertentu dan juga secara spasial. Bunyi spasial pada TAU NIGENS 2020 dihasilkan dengan cara menggerakan alat putar bunyi dari satu posisi ke posisi yang berbeda dan direkam menggunakan *michrophone* khusus yang mampu merekam perubahan arah bunyi yang sumber bunyinya bergerak (Politis dkk., 2020).

Terdapat juga penelitian yang melalukan augmentasi menggunakan data dari DCASE 2019 yaitu penelitian oleh Zhang, dkk (Zhang dkk., 2019). Pada proses augmentasinya dilakukan dengan memanfaatkan penelitian lain yaitu

SpecAugment (Park dkk., 2020). Data augmentasi yang dihasilkan menambah jumlah kelompok data. Aslinya data pada DCASE 2019 memiliki empat pembagian kelompok data, sedangkan pada data augmentasi jumlahnya bertambah sebanyak dua belas kolompok data yang diperoleh dari tiga metode augmentasi SpecAugment. Tiga metode augmentasi SpecAugment yang dilakukan adalah yang pertama dan kedua secara acak menghapus beberapa baris (*frame*) atau kolom (*frequency bins*) sehingga menjadi sebuah data baru. Pada metode ketiga menghapus kedua baris dan kolomnya. Terdapat 200 sampel data untuk masing-masing kelompok (Zhang dkk., 2019). Proses augmentasi ini tidak mengubah jumlah kelas yang digunakan, tapi menghasilkan varian data baru untuk setiap kelas bunyi. Walaupun terdapat kelemahan pada proses augmentasinya dapat menghilangkan informasi sinyal untuk data bunyi yang dihasilkan.

Bab III Metodologi Penelitian

Pada Bab III akan menjabarkan kerangka kerja eksperimen dan data yang digunakan pada eksperimen. Secara umum eksperimen yang dilakukan pada penelitian ini dapat dilihat pada Gambar III.1. Eksperimen dibagi menjadi tiga bagian yaitu eksperimen A, B, C dan D. Eksperimen A merupakan eksperimen awal yang secara tidak langsung mendasari pengembangan algoritma yang diujikan pada eksperimen D, yang menjadi kebaharuan dari algoritma yang dikembangkan. Eksperimen A dilakukan untuk mencari peluang penelitian dari pemisahan bunyi tumpang tindih sehingga proses dari eksperimen A ini menginspirasi pengembangan algoritma penggabungan pemisahan bunyi dengan deteksi dan lokalisasi bunyi yang dilakukan pada eksperimen D. Namun, sebelum dilakukan eksperimen D akan dilakukan eksperiman B dan C terlebih dahulu. Eksperimen C menguji algoritma deteksi dan lokalisasi bunyi saja tanpa adanya proses pemisahan bunyi seperti pada eksperimen D. Hasil eksperimen C dan D akan dibandingkan untuk melihat performa dari algoritma yang dirancang.



Gambar III.1 Pembagian kelompok eksperimen

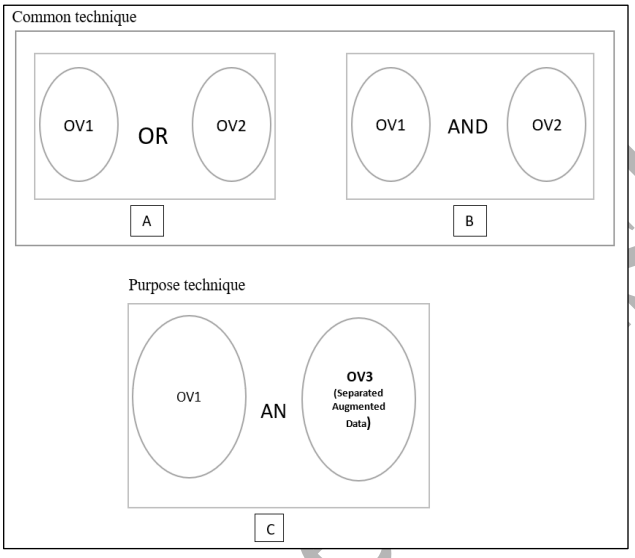
Pembahasan selanjutnya adalah rincian dari masing-masing eksperimen yang dilakukan, diawali dari eksperimen A. Secara rinci eksperimen A ini terbagi menjadi dua yaitu A1 dan A2. Eksperimen A1 menggunakan data tunggal. Pada bunyi tunggal akan dilakukan klasifikasi menggunakan *Support Vector Machine* (SVM). Data yang digunakan pada eksperimen A1 adalah ESC-50, detail dari data ESC-50 ini dijabarkan pada subbab yang berbeda. Eksperimen selanjutnya adalah eksperimen A2 yaitu menggunakan data tumpang tindih yang juga dilakukan klasifikasi dengan SVM. Namun, berbeda dengan eksperimen A1, pada eksperimen

A2 ini akan dilakukan proses pemisahan bunyi tumpang tindih sebelum dilakukan klasfikasi bunyi. Teknik pemisahan bunyi yang dilakukan menggunakan algoritma *Nonnegative Matrix Factorization* (NMF). Data bunyi tumpang tindih yang dilakukan pada eksperimen A2 menggunakan data ESC-50 yang diaugmentasi menjadi data bunyi tumpang tindih. Proses augmentasi akan dijabarkan pada subbab berbeda. Hasil eksperimen A1 dan A2 akan dibandingkan. Hasil eksperimen A dijabarkan pada Bab IV.

Eksperimen dilanjutkan pada eksperimen B. Eksperimen B menggunakan Algoritma NMF sebagai teknik pemisahan bunyi tumpang tindihnya, yang kemudian dilanjutkan dengan melakukan klasifikasi bunyi menggunakan SVM. Hasil dari eksperimen ini akan dibandingkan dengan eksperimen D. Kemudian dilanjutkan dengan eksperimen C melakukan deteksi dan lokalisasi bunyi dengan menggunakan arsitektur SELDnet. Eksperimen C menggunakan dua data yaitu TAU NIGENS 2020 dan TAU NIGENS 2020 yang diaugmentasi. Kemudian eksperimen dilanjutkan dengan eksperimen D yang memiliki tujuan yang sama dengan eksperimen C yaitu deteksi dan lokalisasi bunyi, namun pada eksperimen D dilakukan pemisahan bunyi menggunakan *Time Frequency masking* (T-F *masking*), algoritma yang dikembangkan pada eksperimen D ini disebut dengan Pemisahan SELDnet Data. Hasil dari eksperimen C dan D akan dibandingkan untuk melihat performa dari algoritma yang dikembangkan yaitu Pemisahan SELDnet. Hasilnya dapat dilihat pada Bab IV.

Rancangan kerangka kerja dan eksperimen dibuat berdasarkan kajian teori dari beberapa penelitian terdahulu, kajian ini dapat dilihat pada matriks ringkasan penelitian pada Lampiran B. Selain itu, kerangka kerja penelitian juga dirancang berdasarkan permasalahan bunyi tumpang tindih yang ingin dipecahkan pada penelitian ini. Penggunaan data bunyi tumpang tindih dapat menurunkan performa dari sistem pengenalan bunyi. Permasalahan dengan bunyi tumpang tindih ini telah banyak diteliti dan dipecahkan menggunakan beberapa pendekatan. Berdasarkan jenis data latih yang digunakan, terdapat dua pendekatan yang biasa digunakan untuk pengenalan bunyi tunggal (OV1) dan atau bunyi tumpang tindih (OV2).

Pendekatan yang pertama adalah penggunaan data latih bunyi tunggal untuk pengenalan data bunyi tunggal sedangkan untuk pengenalan bunyi tumpang tindih akan menggunakan data latih bunyi tumpang tindih (Cheong Took, dkk., 2008; Gao, dkk., 2011; Han, dkk., 2015). Pendekatan pertama ini menghasilkan karakteristik data latih yang berbeda antara data tunggal dengan data tumpang tindih, karena pelatihan untuk kedua data dilakukan secara terpisah.



Gambar III.2 Perbandingan Teknik Data Latih

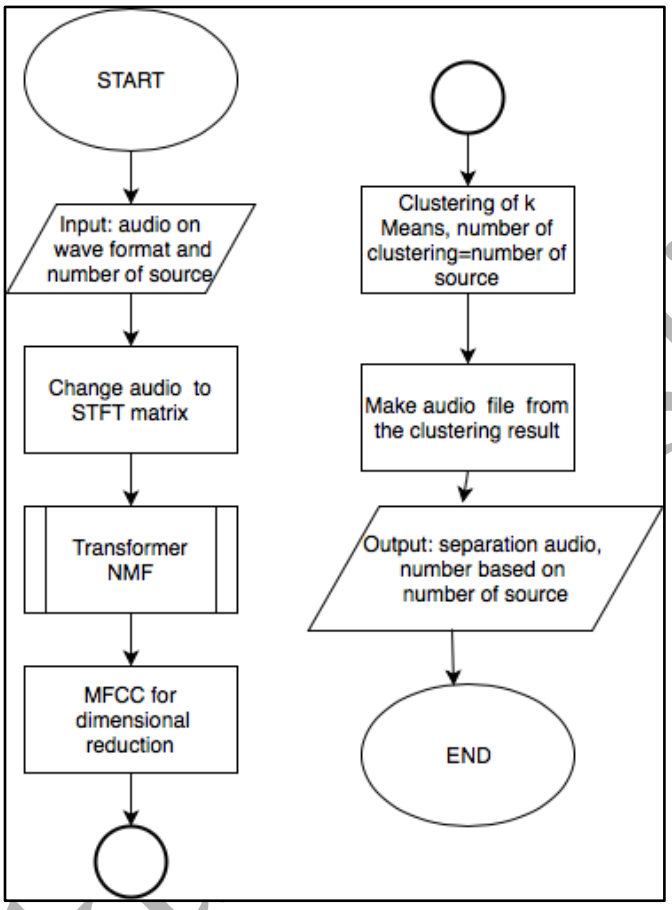
Pendekatan pertama ini dapat disimulasikan pada Gambar III.2 bagian A. Jenis pendekatan kedua adalah data latih yang digunakan berasal dari data tunggal dan data tumpang tindih yang digabungkan untuk kemudian digunakan untuk melakukan pengenalan bunyi dengan kedua kondisi bunyi tunggal dan bunyi tumpang tinidh. Teknik ini memberikan hasil akurasi yang lebih baik untuk data tumpang tindih, tapi untuk dataset tertentu mempengaruhi akurasi data bunyi tunggal menjadi menurun nilai akurasinya (Li, dkk., 2009; Mogi dan Kasai, 2012; Tian, dkk., 2017). Teknik kedua ini dapat dilihat pada Gambar III.2 bagian B. Peningkatan hasil akurasi untuk data bunyi tumpang tindih pada teknik kedua ini menunjukkan adanya peluang untuk meningkatkan performa dari sistem pengenalan bunyi. Namun, oleh karena tingkat akurasi yang juga menurun untuk data tunggal, maka ini menjadi kelemahan pada teknik kedua, untuk itu dikembangkan suatu pendekatan yang berbeda agar tingkat akurasi untuk pengenalan bunyi tumpang tindih meningkat tanpa mempengaruhi tingkat akurasi

untuk bunyi tunggalnya. Pendekatan yang berbeda ini dapat dilihat pada bagian C Gambar III.2. Teknik ketiga inilah yang dikembangkan pada penelitian ini. Pada gambar dapat dilihat bahwa penggunaan data latih adalah data tunggal dan data hasil pemisahan, juga dengan tambahan data sintetis (OV3) untuk memperbanyak varian dari jenis data tumpang tindih. Bagian C ini menjadi bagian kebaharuan dari penelitian yang dilakukan.

III.1 Rancangan Kerangka Kerja *Nonnegative Matrix Factorization* dan *Support Vector Machine* (Eksperimen A)

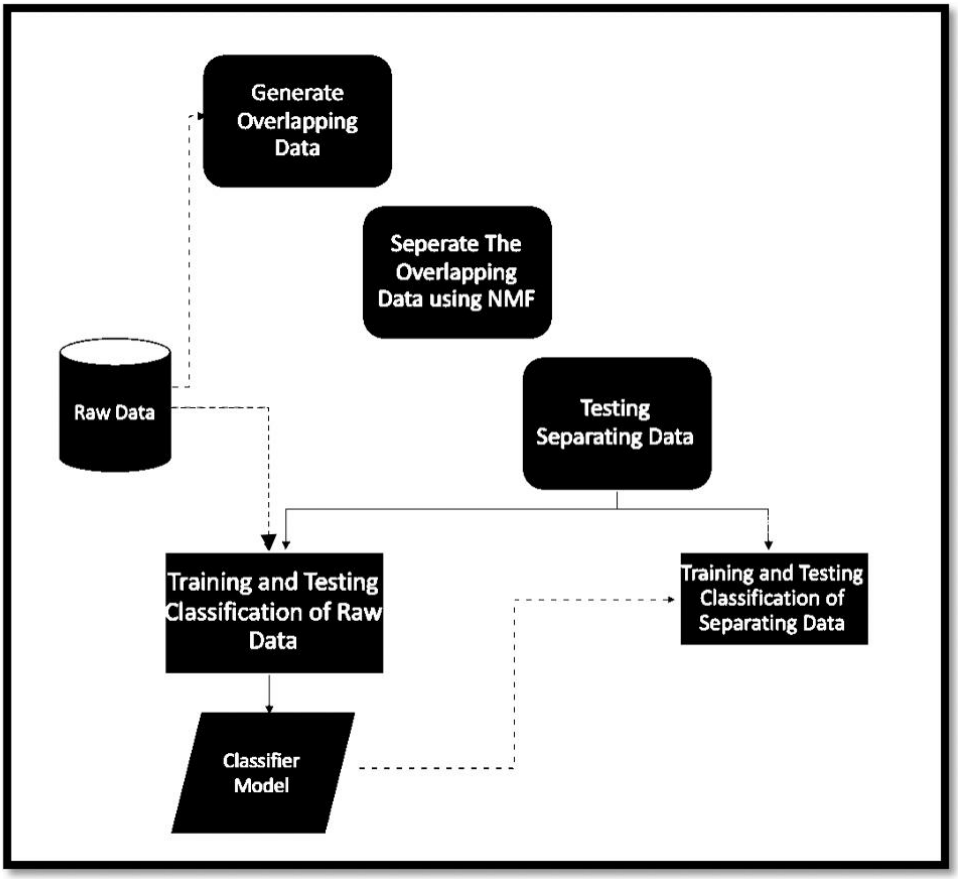
Proses tahapan pemisahan bunyi tumpang tindih menggunakan teknik pemisahan data menggunakan faktorisasi, yaitu *Nonnegative Matrix Factorization*. Proses implentasi pemisahan bunyi tumpang tindih menggunakan NUSSL berbasis bahasa pemrograman *Python* yang dikembangkan oleh Manilow, dkk (2018) dari *Interactive Audio Lab di Massachusetts Institute of Technology* (MIT). Pengkodean NUSSL ini dikembangkan untuk melakukan analisa pada pemrosesan musik sedangkan pada penelitian ini digunakan untuk pengenalan bunyi kegiatan. Pada Gambar III.3 menunjukkan algoritma pengkodean pada NUSSL. Program awalnya menerima input data dengan format *.wav* dan jumlah sumber bunyi yang ingin dihasilkan pada proses pemisahan bunyi tumpang tindih. *File* audio *.wav* ditransformasi menjadi matriks frekuensi dengan menggunakan *Short Time Fourier Transform* (STFT). Pengkodean STFT menggunakan *AudioSignal* berbasis *Python*, penjelasan tentang STFT sendiri terdapat pada subbab II.3.2.1. Setelah nilai STFT didapat dilakukan transformasi NMF. Transformasi NMF dilakukan untuk memfaktorkan matriks input STFT. Hasil dari transformasi NMF adalah dua buah matriks yaitu aktivasi dan *template matrix*. Penjelasan lebih lanjut tentang tahapan NMF dapat dilihat pada subbab II.5.1. Tahapan selanjutnya setelah NMF adalah pengurangan dimensi pada matriks template. Reduksi dimenis menggunakan *Mel Frequency Cepstrum Coefficients* (MFCC), tahapan detail dari MFCC dapat dilihat pada subbab II.3.2. Setelah matriks aktivasi dan *template* didapat maka dilakukan proses *clustering* menggunakan kMeans. Jumlah pemisahan bunyi yang diinput di awal menjadi jumlah kelas yang akan dihasilkan pada tahapan kMeans ini. Kedua matriks akan dilakukan *clustering* secara terpisah di mana nantinya akan dihasilkan

jumlah kelas sesuai input untuk masing-masing matriks. Matriks aktivasi dan template yang telah di-*cluster* akan dipasangkan kembali untuk kemudian diubah kembali menjadi file audio dengan format .wav.



Gambar III.3 *Flowchart* pemisahan bunyi tumpang tindih menggunakan NUSSL berbasis *Python Language*

Hasil pemisahan akan diujikan dengan metode SVM dengan membentuk model klasifikasinya. Model klasifikasi dibentuk dari data yang tidak saling tumpang tindih, model juga diujikan dengan metode SVM sendiri terhadap data tunggalnya. Kerangka kerja dari sistem dapat dilihat pada Gambar III.4.

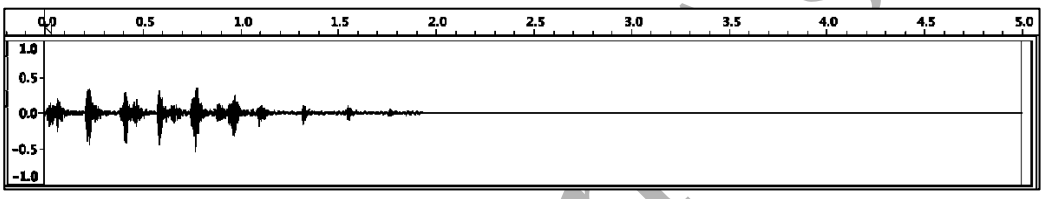


Gambar III.4 Kerangka kerja penelitian

III.1.1 Proses augmentasi data dengan ESC-50

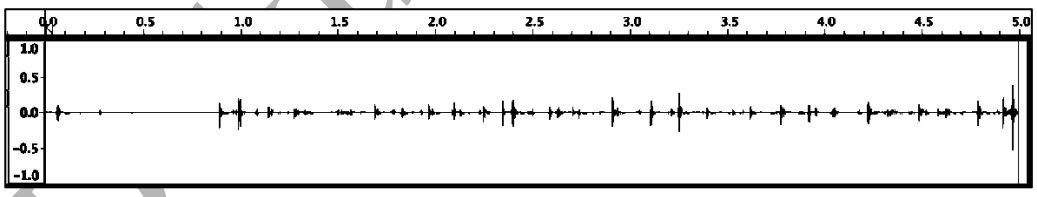
Latar belakang melakukan augmentasi data ESC-50 karena data bunyi yang tersedia adalah data yang bunyi yang saling terpisah jenisnya, sedangkan data yang dibutuhkan adalah data yang saling tumpang tindih atau *overlapping*. Bisa saja data *overlapping* yang dibutuhkan ini dibuat dengan menggunakan *tools editing file* suara dengan format wav atau mp3, tapi oleh karena jumlah data yang banyak dan keterbatasan waktu, maka dibutuhkan *code* yang mampu melakukan penggabungan data secara otomatis. *Code* diambil dari (Robert, 2011) dan (James, 2010). Pada sumber *code*, menggunakan *library pydub (AudioSegment).* Dokumentasi penggabungan lengkap dapat dilihat pada (Robert, 2011). Kemudian *code* dilanjutkan dengan membuat penggabungan data otomatis untuk setiap data wav.

Sumber data yang digunakan untuk membentuk data tumpang tindih adalah data pada penelitian (Piczak, 2015). Pada data tersebut terdapat lima kelas bunyi lingkungan. Masing-masing kelas terdapat 10 jenis bunyi tunggal, dan masing-masing jenis bunyi tunggal tersebut memiliki 40 variasi data wav. Pada penelitian ini hanya akan digunakan dua kelas bunyi yaitu, *human non-speech sounds* dan *interior/domestics sounds*, karena penelitian ini berfokus pada lingkungan di dalam ruangan. Kedua kelas bunyi tersebut masing-masing jenisnya akan dikombinasikan, sehingga dihasilan 400 data tumpang tindih dengan format wav. Berikut adalah contoh penggabungan data yang dilakukan. Data yang digunakan diambil dari ESC-50 yaitu 1-1791-A-26.wav (durasi lima detik, suara orang tertawa), dan 1-137-A-32.wav (durasi lima detik, bunyi ketikan mesin ketik).

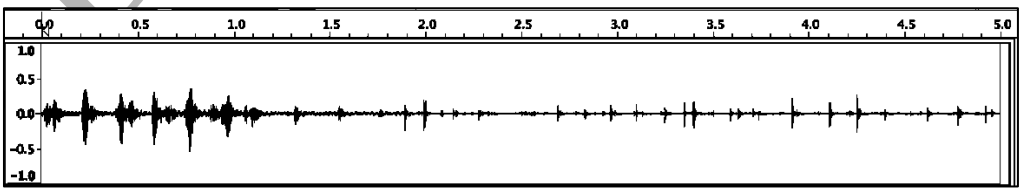


Gambar III.5 Sinyal suara orang tertawa

Gambar III.5 menunjukkan bentuk sinyal suara orang tertawa. Suara orang tertawa masuk pada kategori *human non-speech sounds*. Durasi yang dimiliki file .wav ini adalah lima detik, walaupun sebenarnya suara tertawa hanya sepanjang kuran lebih dua detik, sisanya adalah hening.



Gambar III.6 Sinyal bunyi mesin ketik



Gambar III.7 Sinyal gabungan atau tumpang tindih bunyi ketik dan suara orang tertawa

Gambar III.6 menunjukkan bentuk sinyal bunyi mesin ketik, dengan durasi lima detik. Bunyi yang dihasilkan hampir sepanjang lima detik dengan di antaranya

terdapat hening, berbeda dengan Gambar III.5 yang dimana durasi hening lebih panjang ketimbang durasi bunyinya. Hasil keluaran atau *output* adalah gabungan nama file kedua file dengan format .wav dan durasi 5 detik juga. Bentuk sinyal hasil penggabungan dapat dilihat pada Gambar III.7.

III.2 Rancangan Kerangka Kerja NMF dan SVM (Eksperimen B)

Pada eksperimen B ini akan dilakukan pengujian terhadap algoritma pemisahan NMF dan pengenalan bunyi menggunakan SVM. Pelatihan model pengenalan akan dibangun menggunakan data TAU NIGENS 2020. Hasil model pelatihan akan digunakan untuk melakukan pengenalan bunyi tumpang tindih menggunakan NMF dan SVM. Hasil dari eksperimen ini akan dibandingkan dengan eksperimen D.

III.3 Rancangan Kerangka Kerja Pemisahan SELDnet (Eksperimen C dan D)

Salah satu tujuan dari penelitian yang ingin dicapai adalah meningkatkan akurasi sistem pengenalan bunyi yang tumpang tindih. Berdasarkan penelitian yang telah dilakukan perbedaan akurasi antara bunyi tanpa tumpang tindih dengan bunyi tumpang tindih memiliki perbedaan yang signifikan, bunyi tanpa tumpang tindih memiliki akurasi jauh di atas bunyi dengan tumpang tindih. Pada kondisi nyata, bunyi tumpang tindih sering ditemui dan menjadi tantangan pada proses implementasi sistem bunyi. Berdasarkan permasalah tersebut maka rancangan kerangka kerja tumpang tindih dibuat untuk memecahkan masalah pengenalan bunyi tumpang tindih. Kerangka kerja dikembangkan berdasarkan penelitian sebelumnya yaitu SELDnet yang telah dibahas pada Bab II. Hasil dari SELDnet pada bunyi tumpang tindih memiliki tingkat akurasi yang lebih rendah dari pada bunyi tanpa tumpang tindih. Jadi, untuk meningkatkan akurasi pada deteksi dan lokalisasi bunyi tumpang tindih dikembangkan sebuah kerangka kerja baru yaitu *Separation SELDnet* atau Pemisahan SELDnet. Teknik Pemisahan SELDnet menerapkan proses pemisahan bunyi menggunakan teknik berbasis *time-frequency masking* (T-F). Proses pemisahan T-F ini diawali dengan mengubah nilai *Short Time Fourier Transform* (STFT) menjadi nilai magnitud spektogram. Kemudian, nilai dari magnitud spektogram diproses menjadi nilai *masking* yang akan digunakan pada tahap pemisahan bunyi. Proses pemisahan menggunakan T-F

masking membutuhkan data sampel dari data bunyi tidak tumpang tindih, yang dimana data itu digunakan untuk membentuk data tumpang tindihnya. Bunyi tidak tumpang tindih tersebut digunakan untuk memisahkan bunyi tumpang tindih.

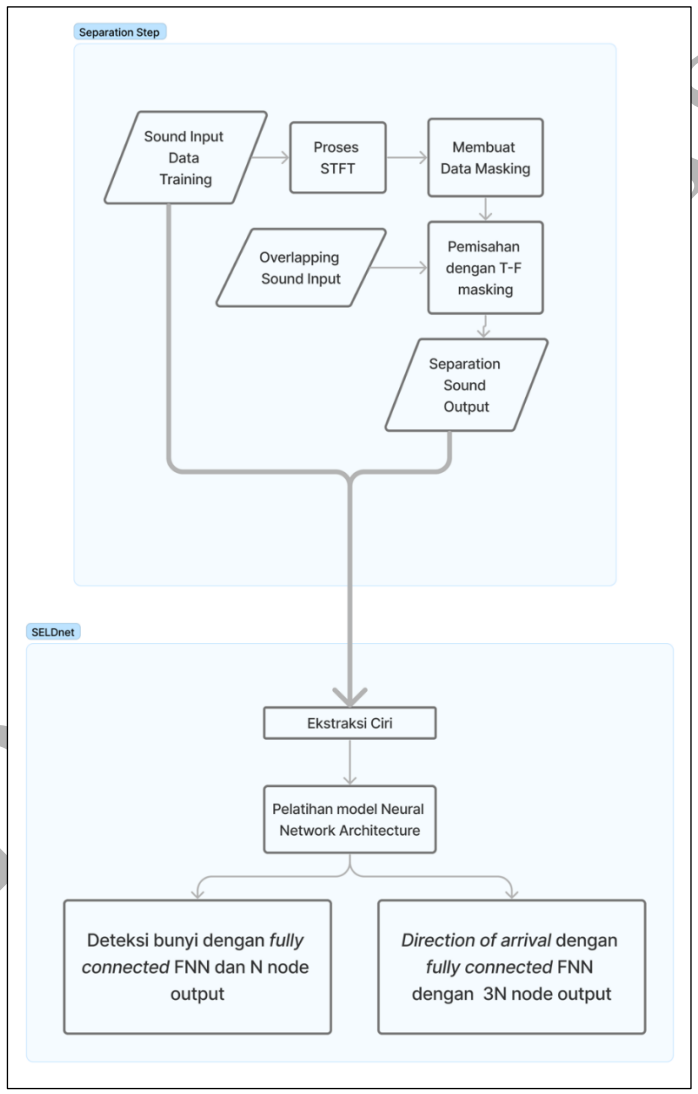
Proses *masking data* merupakan tahap awal dari pemisahan bunyi tumpang tindih. Hasil keluaran proses *masking data* adalah *masked_stft*. Nilai dari *masked_stft* ini diperoleh dari *masked_abs* dan *phase*. Sedangkan nilai *masked_stft* sendiri diperoleh dari perkalian antara nilai magnitude dengan *masked_data*. Nilai absolut dari STFT dataMIX yang menjadi nilai magnitude-nya. Sedangkan nilai *phase* didapat dari STFT bunyi tumpang tindih atau mix. Nilai *mask_data* dihasilkan dengan menghitung data bunyi tunggal dibagi dengan nilai maksimum antara nilai STFT bunyi tunggal dengan nilai STFT dari bunyi mix. Nilai *masked_stft* yang merupakan keluaran dari tahap *masking*, digunakan pada proses pemisahan bunyi dan kemudian hasil dari pemisahan bunyi ini akan di-*inverse* menggunakan inverse STFT, untuk kemudian diubah menjadi format data audio kembali. Gambar III.7 adalah tahapan proses pemisahan menggunakan *Time Frequency masking* dengan

1. *Buat data masking:*
   a. *mask_data = dataSingle/(max(dataMix,dataSingle))+ nussl.constants.EPSILON*
   b. *masked_stft = masked_abs*np.exp(1j*phase)*
      di mana:
      *masked_abs = magnitude*masked_data*
      magnitude = np.abs(dataMix.stft_data)
      phase = np.angle(dataMix.stft_data)
2. *Tahap pemisahan:*
   a. *estSeparation = dataMix.make_copy_with_stft_data(masked_stft)*
   b. *inverse STFT = estSeparation.istft()*
3. *Menghasilkan data audio:*
   estSeparation.write_audio_to_file(folderOutput+ "HasilSeparated.wav", sampe rate=44100)

Gambar III.7 Tahapan *Time Frequency masking*

Hasil dari tahap pemisahan bunyi digunakan sebagai data input pada proses lokalisasi dan deteksi bunyi (SELDnet). Lokalisasi dan deteksi bunyi SELDnet ini diawali dengan ekstraksi ciri, hal ini seperti yang dilakukan pada SELDnet aslinya dan begitu juga proses setelahnya hingga didapat hasil lokalisasi dan deteksi bunyi.

Teknik Pemisahan SELDnet dirancang untuk meningkatkan akurasi dari hasil lokalisasi dan deteksi bunyi. Kerangka kerja dari Pemisahan SELDnet dapat dilihat pada Gambar III. Berdasarkan Gambar III. tersebut dapat dilihat bawah proses pemisahan dilakukan di awal kerangka kerjanya. Jadi, input dari proses SELDnet adalah data tunggal dari hasil pemisahan bunyi. Pada penelitian dengan hanya SELDnet, hasil akurasinya baik saat menggunakan data tunggal, sedangkan pada data tumpang tindih hasil akurasinya masih rendah. Sedangkan pada Pemisahan SELDnet memberikan hasil akurasi yang lebih baik untuk data tumpang tindih dan tidak mempengaruhi hasil akurasi dengan data tunggalnya.



Gambar III.8 Kerangka kerja dari Pemisahan Separation SELDnet

Teknik Pemisahan SELDnet ini diuji performanya menggunakan beberapa dataset. Dataset utama yang digunakan adalah TAU NIGENS *Spatial Sound Events* 2020,

kemudian dari data utama ini dihasilkan data augmentasi lainnya yang memperbanyak variasi dan jumlah dataset untuk data tumpang tindih. Begitu juga untuk teknik SELDnet *only* diuji menggunakan TAU NIGENS 2020 ini. Hasil eksperimen untuk Pemisahan SELDnet dan SELDnet *only* diukur menggunakan *Error rate* dan *F-score*.

III.3.1 Implementasi SELDnet dan Pemisahan SELDnet

Pada tahap implementasi, untuk menentukan performa dari sistem SELDnet ini menggunakan nilai tingkat kesalahan atau *error rate* (ER) serta skor F1 (Adavanne, dkk., 2019; Mesaros, Diment, dkk., 2019). Pada penelitian dengan dua tujuan yang digabungkan ini yaitu deteksi dan lokalisasi, menghasilkan perhitungan performa keduanya secara terpisah, sehingga pengukuran performa menjadi kurang akurat (Mesaros, dkk., 2019). Misal, pada satu pengujian memberikan hasil pengujian yang tepat dengan kondisi sebagai berikut: deteksi memberikan hasil yang tepat berdasarkan label yang telah diberikan, sedangkan hasil lokalisasinya tidak tepat, dan sistem sendiri mengacu pada deteksi karena perhitungan skornya bergantung pada label jenis bunyinya. Hal ini membuat pengukuran menjadi kurang akurat. Masalah pengukuran ini telah dikemukakan pada penelitian yang dilakukan oleh Mesaros, dkk., 2019 dengan solusi yang juga telah diberikan dengan pendekatan pengukuran yang lebih akurat.

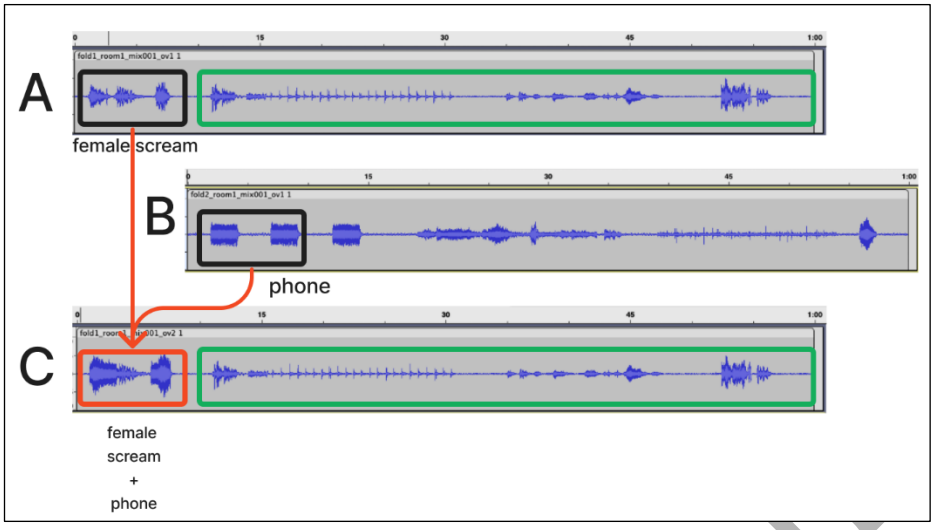
Eksperimen dilakukan untuk mengetahui performa dari SELDnet dan parameter yang mempengaruhi performanya. Pengujian dilakukan dengan dua tujuan yaitu, pertama mencari nilai parameter yang mempengaruhi hasil, *error* dan skor SELDnet, kedua yaitu mencari pengaruh jumlah data dengan performa SELDnet. Pada tujuan eskperimen pertama, terdapat beberapa parameter yang diujikan nilainya, antara lain: jumlah ciri per *frame*, ukuran dari *batch*, serta jumlah *node* pada RNN dan FNN. Perubahan nilai parameter menghasilkan ukuran performa yang berbeda-beda. Salah satu batasan dari perubahan nilai parameter adalah kemampuan mesin, pada nilai tertentu mesin tidak mampu melakukan perhitungan karena keterbatasan memori mesinnya. Sehingga hasil dari uji coba bagian pertama ini akan menentukan nilai parameter yang mempengaruhi, sedangkan nilai

performa terbaik dari SELDnet menggunakan penelitian acuan. Pada data terdapat dua bagian data yaitu data *development* dan *evaluation.* Dengan data *development* diberi label angka 1-6 yang digunakan untuk memisahkan antara pelatihan dan pengujian. Pada eskperimen digunakan label 3-6 untuk pelatihan dan label 2 (dua) untuk validasi dan label 1 (satu) untuk pengujian.

Sebelum melakukan eksperimen pertama dilakukan, dilakukan dulu tahap awal eksperimen. Pada tahap awal eksperimen, akan dilakukan pencarian nilai parameter yang dapat ditangani oleh mesin. Pencarian nilai parameter ini menggunakan mode "quick_test". Mode ini membuat proses SELDnet mulai dari pelatihan, pengujian hingga perhitungan performa, dalam skema yang lebih singkat karena hanya menggunakan sedikit data validasi, yaitu hanya dua buah data validasi dan dua epoch pelatihan. Setelah dilakukan eksperimen "quick_test" dan didapat nilai-nilai parameter yang dapat ditangani oleh mesin, maka dilakukan "full_mode" yang melakukan epoch sebanyak 50 kali dengan nilai parameter dari eksperimen "quick_test".

III.3.2 Augmentasi dataset TAU-NIGENS 2020

Berdasarkan data TAU NIGENS 2020, dilakukan proses augmentasi data. Proses augmentasi data menggunakan data tunggal dari dataset yang ada. Prosesnya diawali dengan memotong bagian tunggal dari dua *file* audio dengan pelabelan ov1, kemudian dari hasil pemotongan ini digabungkan dalam *time frame* yang sama sehingga menghasilkan kondisi bunyi yang tumpang tindih. Simulasi proses augmentasi data dapat dilihat pada Gambar III.9. Berdasarkan gambar dapat dilihat pada Label A dan Label B adalah data bunyi yang tidak tumpang tindih merupakan input dari augmentasi data, kemudian menghasilkan data pada Label C. Hasil augmentasi data ini digunakan untuk menguji teknik Pemisahan SELDnet, yaitu pada eksperimen C.



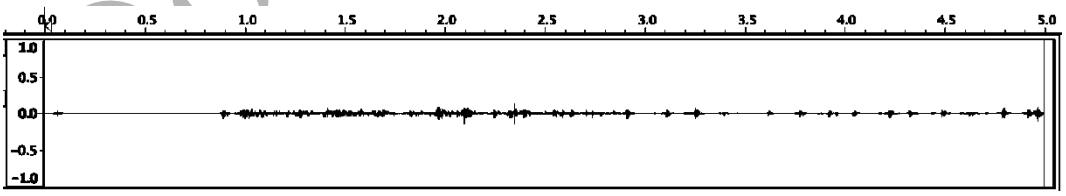
Gambar III.9 Proses Augmentasi Data

Bab IV Analisa Hasil Ekperimen

Hasil eksperimen yang dijabarkan pada Bab IV ini adalah pengujian terhadap rancangan kerangka kerja yang telah disusun pada BAB III. Hasil eksperimen kerangka kerja *Nonnegative Matrix Factorization* (NMF) dan *Support Vector Machine* (SVM) menggunakan data augmentasi dari ESC-50 akan dijabarkan dan diukur akurasinya. Begitu juga hasil eksperimen penerapan kerangka kerja Pemisahan SELDnet juga akan dijabarkan. Namun, sebelum melakukan eksperimen Pemisahan SELDnet, sesuai dengan rancangan eksperimennya, hasil implementasi SELDnet akan dijabarkan terlebih dahulu sebagai pembanding dengan teknik yang dirancang yaitu Pemisahan SELDnet. Kedua hasil ini akan diuji menggunakan data TAU NIGENS 2020 yang juga diagumentasi. Hasil dari eksperimen implementasi SELDnet dan kerangka kerja eksperimen Pemisahan SELDnet akan diukur performanya menggunakan pengukuran skor yang sama yaitu nilai *F-score* dan *Error rate*-nya.

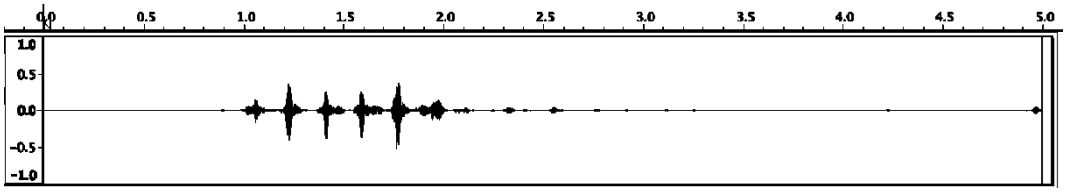
IV.1 Hasil Eksperimen *Nonnegative Matrix Factorization* dan *Support Vector Machine* (Eksperimen A.1 dan A.2)

Berikut adalah simulasi dari pemisahan data tumpang tindih dengan NMF yang diuji dengan proses pengenalan bunyi menggunakan teknik SVM. Input data yang digunakan adalah hasil output atau keluaran dari tahap pembentukan data tumpang tindih. Hasil keluaran terdiri dari dua file yaitu Gambar IV.1 dan Gambar IV.2



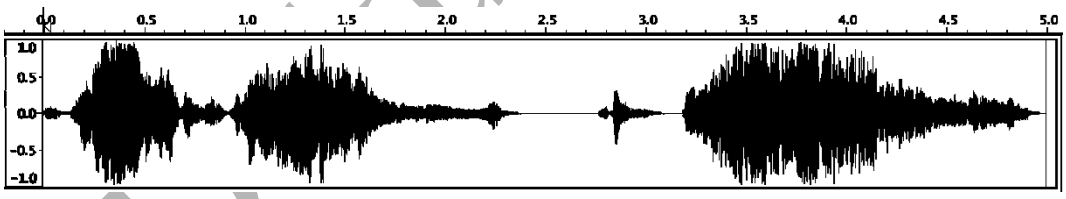
Gambar IV.1 Hasil pemisahan pertama dari bunyi ketik dan suara tertawa

Jika Gambar IV.1 diperdengarkan, maka terdapat bunyi ketik seperti pada Gambar III.6, tapi dengan volume yang lebih kecil. Begitu juga pada Gambar IV.2, jika dibandingkan secara kasat mata, bentuk sinyal pada Gambar IV.2 menyerupai Gambar III.5 yaitu suara orang tertawa.

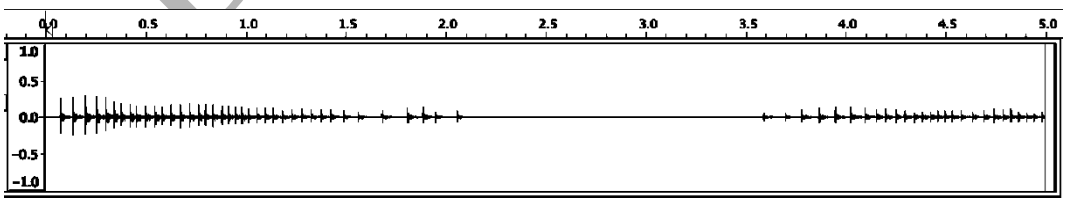


Gambar IV.2 Hasil pemisahan kedua dari bunyi ketik dan suara tertawa

Setelah model klasifikasi dibentuk, model klasifikasi tersebut akan diujikan dengan melakukan klasifikasi data tunggal menggunakan metode SVM. Hasil akurasi klasifikasi data tunggal dilihat pada Tabel IV.1 menunjukkan bahwa hasil akurasi terhadap nilai C, dimana nilai C menunjukkan tingkat *overfitting* dari model klasifikasi yang terbentuk. Semakin tinggi nilai C, maka semakin tingkat *overfitting*-nya. Rata-rata nilai C pada hasil pembentukan model klasifikasi adalah 4 dengan ukuran nilai C minimal 0 dan maksimal 20. Pada Tabel IV.1 rata-rata akurasi klasifikasi untuk seluruh kelas data tunggal adalah 94.135. Akurasi tertinggi diperoleh jenis bunyi *crying_baby* yaitu 99.47% juga dengan nilai C terkecil yaitu 0.6, sedangkan akurasi terendah adalah *door_wood_creaks* dengan akurasi 92.66% dengan nilai C yaitu 5. Jenis bunyi *crying_baby* memang jauh berbeda bentuk sinyalnya dengan rentang amplitudo yang juga berbeda jauh dengan jenis bunyi lainnya. Pada Gambar IV.3 menunjukkan bentuk sinyal *crying_baby*, dan Gambar IV.4 menunjukkan sinyal *door_wood_creaks.*



Gambar IV.3 Sinyal bunyi *crying_baby*



Gambar IV.4 Sinyal bunyi *door_wood_creaks*

Setelah data tumpang tindih berhasil dibuat, tahapan pemisahan dilakukan dengan menggunakan metode NMF pada *nussl project* (Manilow dkk., 2018). Pemisahan

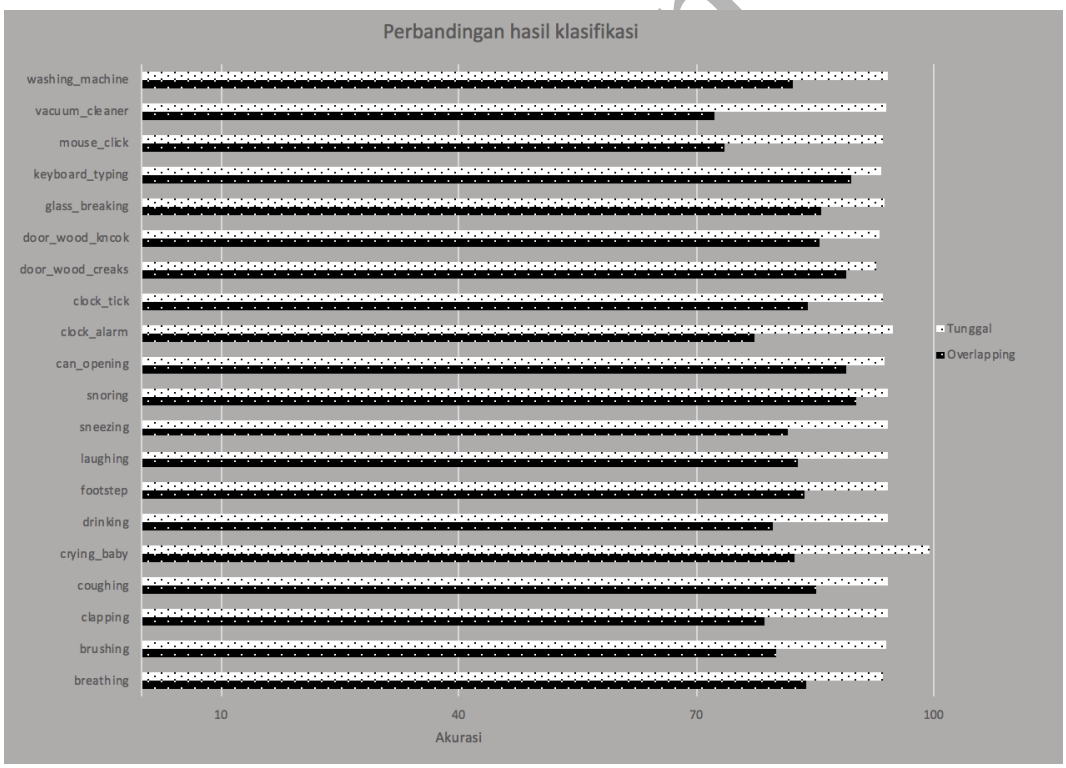
dilakukan terhadap 400 data .wav yang telah dihasilkan. Sehingga akan dihasilkan kembali 80 data .wav untuk dua jenis bunyi yang telah digabungkan, atau dengan kata lain masing-masing jenis terdiri dari 40 data wav. Hasil pemisahan kemudian diklasifikasikan dengan mencocokan berdasarkan jenis bunyi gabungannya. Jadi, misalkan dua jenis bunyi yaitu, tepuk tangan dan gelas pecah digabungkan, artinya terdapat 40 data .wav yang dihasilkan, maka seharusnya dihasilkan data tepuk tangan sebanyak 40 dan gelas pecah sebanyak 40 juga. Hasil 80 data pemisahan tersebut diujikan dengan melakukan klasifikasi SVM dengan model yang sudah dibuat sebelumnya pada tahap pelatihan data tunggal. Jika dari 80 data dikenali 40 data untuk masing-masing kelas maka akurasi yang dicapai adalah 100 persen. Rata-rata hasil akurasi untuk data tumpang tindih adalah 83%. Hasil akurasi klasifikasi data tumpang tindih dapat dilihat pada Tabel IV.1. Hasil pemisahan terbaik berdasarkan rata-rata akurasinya adalah jenis bunyi *snoring* dan hasil pemisahan terburuk adalah *clapping*.

Tabel IV.1 Hasil Akurasi Data Tumpang Tindih

Jenis bunyi	Akurasi	C
door_wood_creaks	92.66	5.0
door_wood_kncok	93.03	4.9
keyboard_typing	93.24	3.9
breathing	93.47	9.2
clock_tick	93.51	5.0
mouse_click	93.52	4.0
glass_breaking	93.67	4.2
can_opening	93.73	7.8
vacuum_cleaner	93.98	4.0
brushing	94.01	4.1
clapping	94.08	4.1
footstep	94.10	3.6
coughing	94.12	3.8
drinking	94.14	3.8
laughing	94.17	3.5

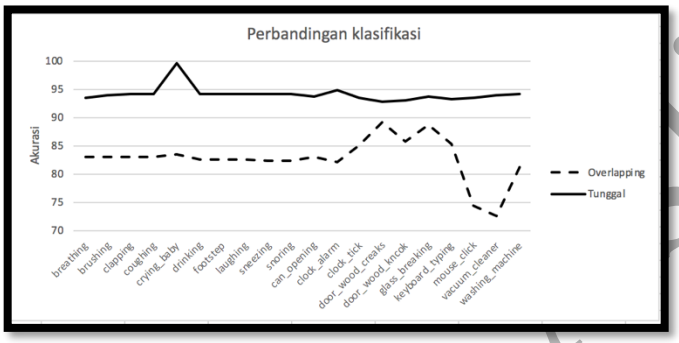
Jenis bunyi	Akurasi	C
sneezing	94.19	3.5
washing_machine	94.22	3.8
snooring	94.23	3.3
clock_alarm	94.80	6.4
crying_baby	99.47	0.6

Pada Gambar IV.5 menunjukkan perbandingan hasil klasifikasi antara data tunggal dengan hasil pemisahan data tumpang tindih. Pada grafik menunjukkan bahwa hasil akurasi tertinggi pada data tunggal tidak sama seperti data tumpang tindihnya. Pada data tunggal akurasi terbaik ada pada *crying_baby*, sedangkan pada data tumpang tindih ada pada jenis *snoring.* Perbandingan akurasi antara data tumpang tindih dengan data tunggal dapat dilihat pada Gambar IV.6.



Gambar IV.5 Perbandingan akurasi klasifikasi antara data tunggal dengan data tumpang tindih

Berdasarkan hasil akurasi yang diperoleh terlihat jika hasil akurasi pada data tumpang tindih masih lebih rendah jika dengan hasil data tunggalnya. Pada klasifikasi data tunggal hasil akurasi cenderung stabil. Pada data tumpang tindih, data bunyi yang berasal dari suara manusia cenderung stabil, sedangkan pada data lingkungan yang bukan berasal dari suara manusia menghasilkan tingkat akurasi yang signifikan berbeda satu dengan yang lainnya.



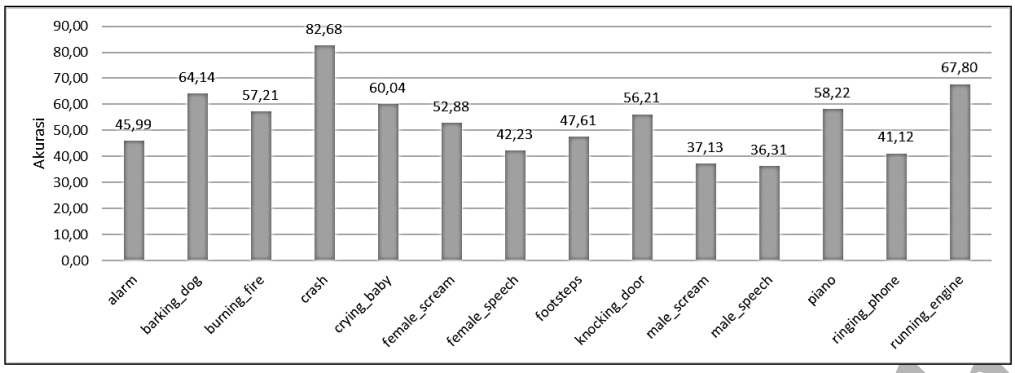
Gambar IV.6 Grafik perbandingan klasifikasi antara data overlapping dengan data tunggal

Berdasarkan hasil uji coba dan analisa yang telah dilakukan, metode NMF telah memberikan hasil pemisahan bunyi yang cukup signifikan, walaupun masih lebih rendah dari pengenalan data tunggalnya. Hal ini menandakan bahwa pemisahan bunyi yang dilakukan belum mencapai tingkat pemisahan yang sama seperti data tunggal awalnya. Namun, telah berhasil melakukan pemisahan dengan cukup baik jika dilihat dari rata-rata akurasi klasifikasi yang dicapai. Penelitian yang dilakukan dapat dikembangkan dengan mengembangkan metode pemisahan yang dapat menghasilkan pemisahan data yang lebih baik lagi serta dengan tidak bergantung pada variasi data bunyi lingkungan yang dilakukan.

IV.2 Hasil eksperimen NMF dan SVM (Eksperimen B)

Berdasarkan rancangan eksperimen B pengujian dilakukan menggunakan data TAU NIGENS 2020 dan Augmentasi TAU NIGENS 2020. Hasil eksperimen dapat dilihat pada Gambar IV.7. Pada Gambar IV.7 menunjukkan bahwa hasil F_score terbaik diperoleh sebesar 82,68 persen pada kelompok data *crash*. Rata-rata akurasi dari eksperimen B ini adalah 53,54 persen. Hasil dari eksperimen B ini akan

dibandingkan dengan eksperimen C dan D yang akan ditampilkan hasil perbandingannya pada subbab berikutnya.



Gambar IV.7 Hasil Akurasi Eksperimen B

IV.3 Hasil eksperimen SELDnet dan Pemisahan SELDnet (Eksperimen C dan D)

Sesuai dengan rancangan eksperimennya, untuk menguji Pemisahan SELDnet ini, dilakukan dua eksperimen yaitu eksperimen SELDnet dan eksperimen Pemisahan SELDnet. Pada eksperimen SELDnet sendiri juga dilakukan pra eksperimen untuk mencari nilai parameter yang dapat ditangani oleh mesin yang kemudian nilai ini akan digunakan untuk eksperimen SELDnet dan Pemisahan SELDnet dengan "full_mode" yaitu dengan 50 epoch. Pada hasil eksperimen "quick_test" dapat dilihat nilai parameter yang tidak dapat ditangani oleh mesin, bahkan pada epoch pertamanya.

Tabel IV.**2** menunjukkan beberapa kondisi nilai parameter yang tidak dapat ditangani oleh mesin. Seperti pada index er0-er5 dengan status "Out of memory", artinya pada kondisi nilai parameter tersebut epoch pertama pun tidak dapat mengeluarkan hasil karena keterbatasan mesin. Nilai-nilai parameter yang dapat ditangani oleh mesin menjadi acuan pada eskperimen pertama ini.

Tabel IV.2 Perubahan nilai parameter untuk mengukur kemampuan mesin

Index	label_sequence_length	batch_size	nb_cnn2d_filt	rnn_size	jml layers	fnn_size	nb_epochs	Status
er0	**64**	**256**	**64**	**128,128**	**2**	**128**	**50**	**Out of memory**
1	64	128	64	128,128	2	128	50	
2	64	128	64	64,64	2	64	50	
er1	**64**	**128**	**128**	**64,64**	**2**	**64**	**50**	**Out of memory**
3	64	128	64	256,256	2	256	50	
4	64	128	64	256,256	2	128	50	
5	64	128	64	**512,512**	2	**512**	50	
er2	**128**	**128**	**64**	**64,64**	**2**	**64**	**50**	**Out of memory**
6	128	32	64	64,64	2	64	50	
7	256	32	64	64,64	2	64	50	
8	256	64	64	64,64	2	64	50	
er3	**512**	**64**	**64**	**64,64**	**2**	**64**	**50**	**Out of memory**
9	256	64	**64**	128,128	2	128	50	
er4	256	128	**64**	128,128	2	128	50	**Out of memory**
er5	256	128	**64**	**64,64**	**2**	**64**	50	**Out of memory**
10	256	64	32	128,128	2	128	50	
11	256	64	32	64,64,64	3	64	50	
12	256	64	64	64,64,64	3	64	50	
13	256	64	64	64,64,64	3	64	70	

Tabel IV.3 Parameter pada eksperimen *full mode*

Index	label_sequence_length	batch_size	nb_cnn2d_filt	rnn_size	jml layers	fnn_size	nb_epochs
A	64	128	64	256	2	256	50
B	128	64	64	128	3	128	50
C	64	64	128	128	2	128	50

Pada

Tabel IV.**2** menunjukkan perubahan nilai parameter. Parameter yang digunakan adalah *label_seqeunce_length*, *batch_size*, *nb_cnn2d_filt*, *rnn_size*, *fnn_size*, *nb_epochs*. Paramater *label_sequence_length* merupakan parameter yang mengatur nilai panjang ciri yang digunakan pada tahap CNN dengan variasi nilai yang digunakan adalah 64, 128, 256, dan 512. Paramater maksimum yang membatasi mesin adalah 512, walaupun dengan nilai *batch_size* kecil. Paramater *batch_size* adalah jumlah data yang digunakan untuk setiap pelatihan dalam pembentukan jaringan CNN dengan variasi nilai: 32, 64, 128, dan 256. Nilai *batch_size* yang menjadi batasan mesin adalah 256 dengan jumlah panjang ciri terkecil. Perubahan nilai RNN dan FNN, baik itu jumlah nodenya maupun jumlah layer tidak berpengaruh banyak terhadap kemampuan mesin. Pada parameter "*nb_cnn2d_filt*" nilai batasan mesin adalah 128, artinya hanya nilai 64 dan 32 yang dapat digunakan untuk menentukan jumlah node pada jaringan CNN. Jumlah epochs pada mode "quick_test" adalah 2 (dua), sehingga hasil performanya tidak menunjukkan hasil

sebenarnya dari sistem SELDnet, karena tidak mengambil model terbaik dengan epoch yang lebih besar.

Hasil batasan nilai parameter dari percobaan awal ini akan digunakan pada percobaan selanjutnya. Percobaan berikutnya dilakukan dengan *full mode*, artinya pencarian model arsitektur *Neural Network*-nya dilakukan sebanyak 50 epochs, untuk kemudian digunakan model terbaiknya pada tahap validasi maupun *testing*-nya. Tabel IV.3 menunjukkan tiga kondisi parameter yang digunakan pada eksperimen *full mode* ini. Setelah didapat parameter yang dapat ditangani oleh mesin, selanjutnya dilakukan eksperimen dengan "full_mode*"* menggunakan nilai parameter pada Tabel IV.3.

IV.3.1 Hasil Eksperimen SELDnet

Hasil percobaan pada parameter terpilih pada Tabel IV.3 ditunjukkan pada Tabel IV.4. Kolom *Index* pada Tabel IV.3 sama dengan kolom *Index* pada Tabel IV.4, artinya nilai parameter pada Index Tabel IV.3 memiliki hasil eksperimen *full mode* pada Index Tabel IV.4 yang sama. Kolom *Split Mode* menunjukkan untuk masing-masing parameter terbagi menjadi lima bagian percobaan berdasarkan kelompok data latih dan uji yang digunakan. Kelompok data telah diberikan label 1, 2, 3, 4, 5, 6. Pada mode *validation* menggunakan data 3, 4, 5, 6 sebagai kelompok data latih dan kelompok data 2 digunakan untuk validasi. Sedangkan mode *test* menggunakan kelompok data 3, 4, 5, 6 sebagai kelompok data latih dan kelompok data 1 sebagai data ujinya. Pengujian juga dilakukan terpisah untuk data dengan kondisi tanpa data tumpang tindih (OV1) dan data dengan kondisi tumpang tindih (OV2). Selain kondisi data bunyi yang tumpang tindih, juga digunakan data uji dengan kondisi gema yang telah ditambahkan secara manual (IR).

Tabel IV.4 Hasil eksperimen *full mode*

	Class aware localization scores									
Score	DOA_error					F_score				
Split / Mode / Index (best epoch)	Validation	Test	Ov1	Ov2	IR1	Validation	Test	Ov1	Ov2	IR1
A (36)	23	24.1	17.8	27.9	23.5	64.3	60.3	70.5	54.3	60.4
B (43)	22.1	22.2	16.3	25.3	21.5	63.5	69.6	68.6	54.4	59.6
C (43)	21	22.4	17.1	25.1	21.7	65.5	61.7	70.1	56.9	61.7
	Location-aware detection scores									
Score	Error Rate					F_score				
Split / Mode / Index (best epoch)	Validation	Test	Ov1	Ov2	IR1	Validation	Test	Ov1	Ov2	IR1
A (36)	0.69	0.72	0.61	0.78	0.72	39.7	36.7	49.8	29.5	37.1
B (43)	0.51	0.71	0.6	0.75	0.7	40.1	38.9	49.8	33.3	39.4
C (43)	0.66	0.71	0.59	0.76	0.7	44.4	39.4	52.7	32.6	40
Score	**seld_score**									
Split / Mode / Index (best epoch)	Validation		Test		Ov1		Ov2		IR1	
A (36)	0.44		0.47		0.38		0.52		0.47	
B (43)	0.44		0.46		0.38		0.5		0.46	
C (43)	0.42		0.45		0.37		0.5		0.45	

Pada kolom *Class aware localization* terbagi menjadi dua bagian kolom yaitu *DOA_error* dan *F_score.* Perhitungan *DOA_error* dapat dilihat pada rumus (4) dan perhitungan *F_score* merujuk pada rumus (1). Hasil *DOA_error* terbaik didapat pada *Split mode Validation* yang tidak berbeda jauh hasilnya dengan *Split mode Test*, selisih 0.87, yang artinya SELDnet untuk proses lokalisasi bunyi berdasarkan *Azimuth* dan *Elevation degree*-nya dengan memperhatikan kelas bunyinya, memiliki performa yang tidak jauh berbeda untuk data uji yang sama maupun tidak sama dengan data latihnya. Kondisi *Split mode OV1* jauh lebih baik dibanding *Split mode OV2* dengan selisih 9.03, yang artinya SELDnet lebih cocok untuk data yang tidak saling tumpang tindih, untuk data yang saling tumpang tindih masih memiliki tingkat *error* yang lebih tinggi. Namun, untuk data *Split mode IR* memiliki hasil yang lebih buruk dibanding OV1, tapi lebih unggul dibanding OV2, menunjukkan performa SELDnet untuk kondisi data gema dapat lebih baik dibanding data tumpang tindih, walaupun masih lebih buruk tanpa tumpang tindih dan tanpa gema. Secara rata-rata keseluruhan hasil eksperimen untuk lokalisasi bunyi dengan memperhatikan kelas bunyinya adalah 22.06 derajat, dimana nilai *DOA_error* yang diharapkan adalah nol, maka performa SELDnet masih dapat ditingkatkan agar semakin mendekati nilai nol untuk berbagai kondisi data. Hasil *DOA_error* berdasarkan parameter yang digunakan, menunjukkan performa yang berbeda untuk setiap *Split mode*. Misal, pada *Split mode Validation* yang memiliki *error* terendah adalah pada parameter *Index C*, dimana parameter pada *Index C* memiliki jumlah node CNN terbanyak dibanding *Index* lainnya. Namun, jika dibandingkan

perubahan error antara *Index A* dengan *Index B* yang menurun sebanyak 0.9 dengan *Index B* dengan *Index C* sebanyak 1.1. Sehingga performa SELDnet dipengaruhi oleh panjang sekuen ciri dan banyaknya node pada CNN adalah hampir seimbang. Hal ini juga berlaku untuk *Split mode OV2* dimana *error* terendah DOA-nya terdapat pada parameter *Index C.* Pada *split mode test, OV1* dan *OV2 error* terendahnya ada pada *Index B,* walaupun selisihnya tidak berbeda signifikan dengan *Index C*. Masih pada bagian *Class aware localization* untuk pengukuran *F_score* nilai yang diharapkan adalah 100. Rata-rata terbaik nilai *F_score* diperoleh pada OV1 yaitu sebesar 69.73 persen, artinya SELDnet mampu melakukan lokalisasi bunyi dengan kondisi tanpa tumpang tindih dan tanpa kondisi gema. Sedangkan rata-rata terendahnya diperoleh pada OV2, sebesar 55.2 yang artinya hanya lebih sedikit dari separuh hasil pengujian yang tepat hasil lokalisasinya untuk data yang saling tumpang tindih, SELDnet untuk data tumpang tindih masih lebih jauh dari yang diharapkan dibanding data tanpa tumpang tindih, begitu juga dengan kondisi data gema yang memperoleh *F_score* sebesar 60.57. Secara keseluruhan nilai *F_score* masih dibawah nilai 100 yang diharapkan, sehingga dapat dilakukan peningkatan. Jika mengacu pada perubahan parameter yang dilakukan, *F_score* terbaik didapat pada *Index C*, selain itu perubahan nilai *F_score* untuk masing-masing parameter lebih berbeda siginifikan dibandingkan dengan perubahan nilai *DOA_error*.

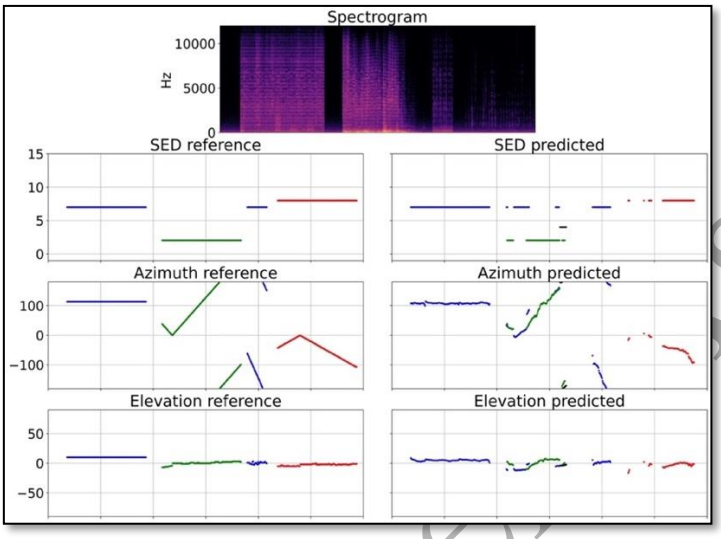
Analisa hasil selanjutnya adalah kolom *Localization-aware detection scores* yang terbagi juga terbagi menjadi dua bagian, yaitu *Error rate* dan *F_score*. Hasil *Error rate* yang diharapkan adalah nol, sehingga nilai rata-rata *Error rate* terbaik ada pada data tanpa tumpang tindih atau OV1 sebesar 0.60. Jika dilihat pada perubahan parameter yang diujikan, nilai *Error rate* antara *Index B* dan *C* tidak berbeda signifikan, bahkan ada yang bernilai sama untuk masing-masing *Split mode*-nya. Hal ini menandakan bahwa perubahan nilai parameter panjang sekuen ciri dan banyaknya node CNN memberi pengaruh yang sama terhadap performa SELDnet. Bagian kedua dari *Location-aware detection scores* adalah *F_score* yang juga mengukur performa SELDnet untuk setiap kondisi *Split mode* dengan tiga jenis perubahan parameter. Sama seperti nilai rata-rata terbaik *Error rate*, nilai terbaik

F_score juga diperoleh pada *Split mode* OV1, yaitu data tanpa tumpang tindih dan kondisi gema yang ditambahkan. Sedangkan untuk perubahan parameter *F_score* nilai terbaiknya untuk setiap *Split mode* didominasi pada perubahan parameter *Index C.* Sehingga performa SELDnet untuk *F_score*-nya lebih dipengaruhi oleh banyaknya node CNN yang digunakan pada proses pelatihan dan pengujiannya. Bagian selanjutnya yang akan dibahas adalah *seld_score.* Nilai *seld_score* yang diharapkan adalah nol untuk performa terbaik SELDnet. Nilai terbaik *seld_score* berdasarkan *Split mode*-nya adalah data tanpa tumpang tindih, yaitu sebesar 0.38. Sedangkan untuk perubahan nilai parameter, nilai *seld_score* terbaik ada pada *Index C*, artinya untuk pengaruh performa SELDnet pada pengukuran *seld_score* ada pada jumlah node CNN yang digunakan.

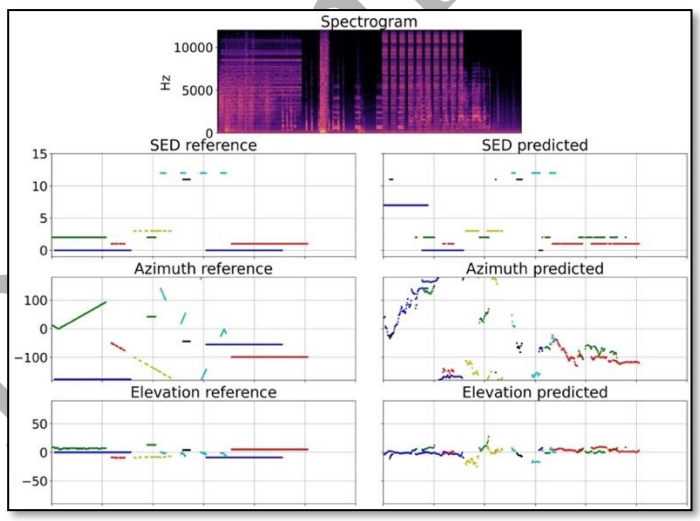
Setelah menganalisa hasil berdasarkan skor, analisa selanjutnya menggunakan grafik perbandingan antara nilai rujukan atau *reference* dengan hasil prediksi atau *predicted.* Pada analisa grafik perbandingan akan digunakan mode validasi tidak terlihat atau *unseen validation* tanpa tumpang tindih. Salah satu hasil validasi tidak terlihat (*unseen validation*) menggunakan data dengan kondisi tanpa tumpang tindih dapat dilihat pada Gambar IV.8. Paramater yang digunakan adalah parameter *Index C.* Pada bagian atas visualisasi berupa *spectrogram* dari data yang diujikan. Bagian kiri, *SED reference, Azimuth reference,* dan *Elevation reference* adalah nilai rujukan yang didapat dari pelatihan, sedangkan pada bagian kanannya adalah hasil prediksi pengujian. Setiap warna garis mewakili kelas jenis bunyi. Garis horizontal sumbu x pada seluruh bagan adalah *time frame* dari data bunyi lingkungan. Sedangkan garis vertikal sumbu y pada SED *reference* dan *predicted* menunjukan index kelas jenis bunyi lingkungan nilainya dari 0-13. Garis vertikal sumbu y pada *Azimuth* dan *Elevation reference* serta *predicted* adalah derajat nilai *Azimuth* dan *Elevataion* untuk setiap jenis bunyi. Semakin mirip letak dan bentuk garis antara *reference* dengan *predicted* artinya semakin baik performa dari SELDnet terhadap data uji yang terpisah dari data latihnya.

Pengujian juga menggunakan data yang saling tumpang tindih. Hasil dari pengujian dengan data tumpang tindih dapat dilihat Gambar IV.8. Jika dibandingkan antara

data bunyi tanpa tumpang tindih dengan data bunyi tumpang tindih seperti pada Gambar IV.9, maka hasil dengan tumpang tindih tidak lebih baik. Hal ini menandakan SELDnet belum cukup tangguh untuk menangani kondisi data yang saling tumpang tindih. Walaupun sudah melakukan pendekatan yang menggabungkan data latih antara data tanpa tumpang tindih dengan data yang tumpang tindih.



Gambar IV.8 Visualisasi hasil keluaran SELDnet tanpa tumpang tindih

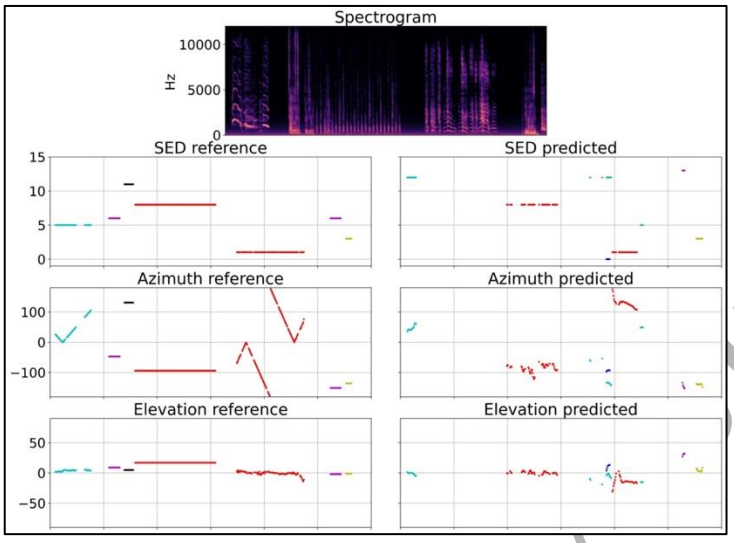


Gambar IV.9 Visualisasi hasil keluaran SELDnet dengan tumpang tindih

IV.3.2 Hasil eksperimen Pemisahan SELDnet

Setelah dilakukan eksperimen dengan teknik SELDnet, selanjutnya adalah mengukur performa dari teknik yang dirancang, yaitu Pemisahan SELDnet. Pada awalnya dilakukan eksperimen pendahulu yang menggunakan data asli TAU

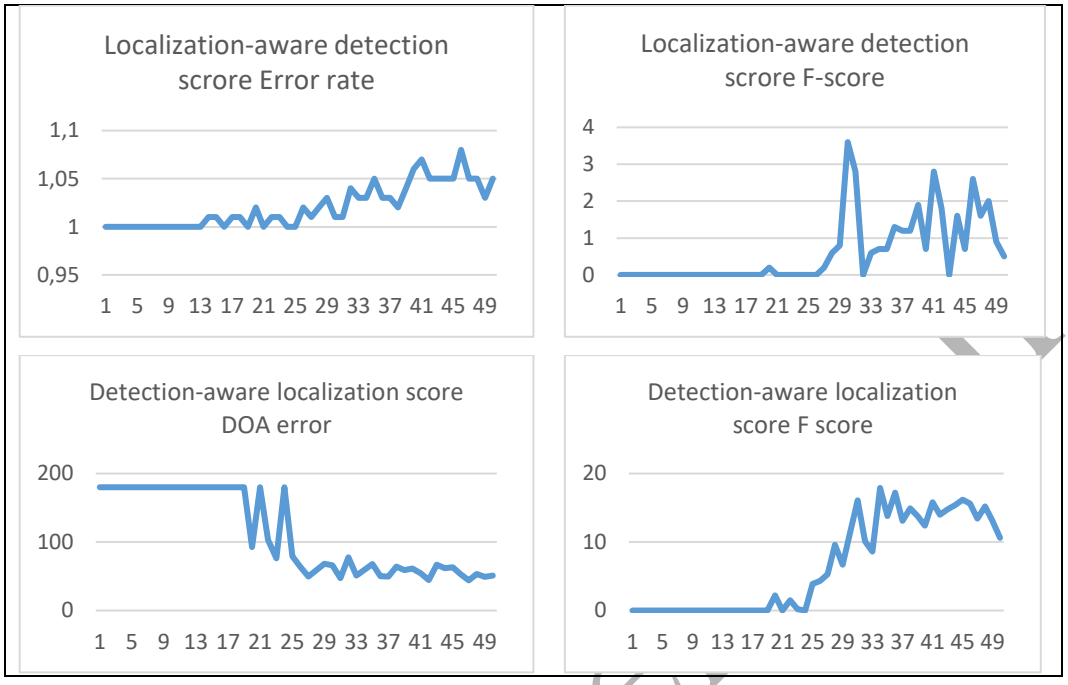
NIGENS 2020 yang kondisi datanya tumpang tindih atau dilabelkan dengan "ov2". Namun, hasil dari eksperimen ini tidak menunjukkan performa Pemisahan SELDnet bekerja dengan baik.



Gambar IV.10 Grafik hasil eksperimen SELDnet dengan data TAU NIGENS 2020

Pada Gambar IV.10 menunjukkan salah hasil pra eksperimen Pemisahan SELDnet dengan data asli tumpang tindih TAU NIGENS 2020. Visualisasi hasil menggunakan teknik yang sama seperti eksperimen sebelumnya yaitu eksperimen SELDnet dengan data asli TAU NIGENS. Berdasarkan Gambar IV.10 dapat dilihat bahwa hasil deteksi (SED) dan lokalisasinya berdasarkan label *Azimuth* dan *Elevation*-nya masih belum memiliki performa yang baik karena bentuk grafik antara *reference* dengan *predicted*-nya masih berbeda jauh. Hal ini terjadi karena proses validasi pemisahan tidak akurat. Setelah proses pemisahan pada *file* audio tumpang tindih menggunakan T-F *masking*, kemudian dilakukan proses identitifikasi dan lokalisasi. Pada saat proses identifikasi dan lokalisasi ini dibutuhkan metadata dari *file audio* untuk pelatihan model pengenalan. Input dari metadata juga sudah disesuaikan dengan kondisi keluaran proses pemisahan bunyi. Namun, setelah dicek kembali, ternyata banyak terjadi kesalahan *labeling*, oleh karena hasil pemisahan tidak sesuai dengan label yang seharusnya, karena proses pemisahan label dilakukan secara manual. Berdasarkan hasil pra eksperimen dengan data asli tumpang tindih ini, maka diputuskan untuk

melakukan proses augmentasi data dengan tujuan dapat mengurangi kesalahan label jenis bunyi hasil pemisahan. Proses augmentasi ini sendiri sudah dijelaskan pada subbab sebelumnya.



Gambar IV.11 Detail hasil eskperimen Pemisahan SELDnet dengan TAU NIGENS 2020

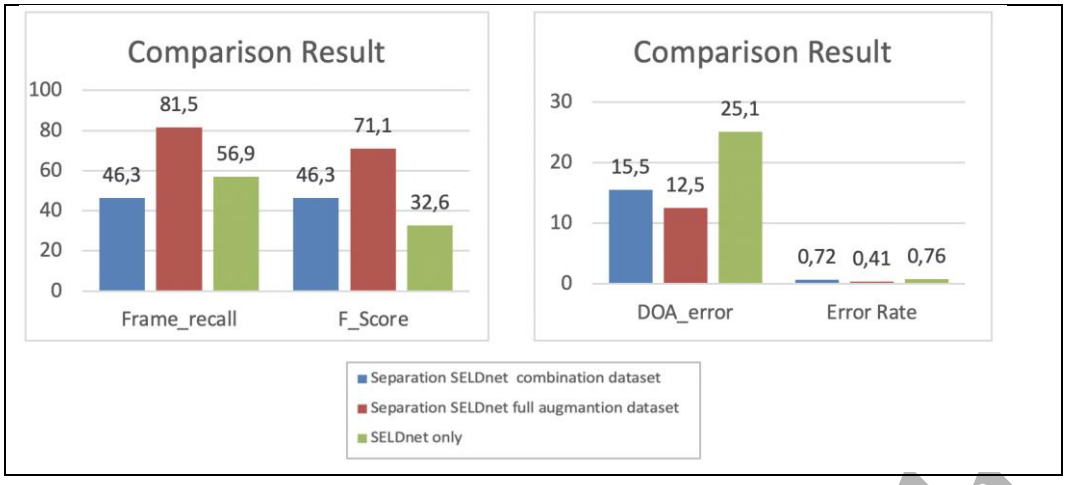
Begitu juga jika dilihat dari hasil keseluruhan nilai validasi berdasarkann *localization-aware detection score* dan *detection-aware localization score* masih belum menunjukkan hasil performa yang baik. Detail dari hasil pra eksperimen ini dapat dilihat pada Gambar IV.11, yang memperlihatkan detail hasil percobaan per epoch-nya, jumlah epoch sebanyak 50. Dapat dilihat juga bahwa peningkatan performa terjadi mulai pada epoch di atas epoch ke-10, walaupun tidak stabil peningkatannya. Pada epoch tertentu malah menunjukan penurunan performa, hal ini juga mengindikasikan pelatihan model yang dibentuk tidak cukup stabil terhadap data. karena akurasi maksimum yang berhasil dicapai sangat jauh dibawah hasil akurasi *benchmark*-nya, yaitu skor untuk deteksi dibawah 5 persen untuk *F-score*-nya. Begitu juga nilai *error rate*-nya sangat tinggi yaitu di atas 0,5. Skor untuk lokalisasi juga rendah yaitu 5 persen untuk *Frame recall*-nya dan untuk *error rate* skornya 180 derajat.

Selanjutnya dilakukan eksperimen menggunakan data augmentasi. Berdasarkan penggunaan data latihnya, eksperimen Pemisahan SELDnet dengan data augmentasi TAU NIGENS 2020 sendiri terbagi menjadi tiga kelompok. Tujuannya agar hasil eksperimen dapat membandingkan performa antara teknik dasar SELDnet tanpa pemisahan dengan teknik yang dirancang yaitu Pemisahan SELDnet. Hasil pengujian terhadap kelompok data yang digunakan sesuai dengan rancangan kebaharuan yang juga dapat dilihat pada bagian ini. Kelompok data latih yang digunakan dibagi berdasarkan jenis data asli yaitu:

1. Kelompok pertama adalah data augmentasi TAU NIGENS 2020 yang telah dipisahkan dengan teknik Pemisahan SELDnet dikombinasikan dengan data asli tunggal TAU NIGENS 2020. Kelompok pertama ini akan disebut sebagai *Separation SELDnet combination dataset*.
2. Kelompok kedua merupakan penggunaan keseluruhan data latihnya adalah data pemisahan data tumpang tindih dari augmentasi data tunggal TAU NIGENS 2020. Kelompok kedua ini akan disebut sebagai *Separation SELDnet full augmentation dataset.*
3. Kelompok ketiga adalah data asli TAU NIGENS 2020 dengan kondisi tunggal dan tumpang tindah. Pada kelompok data ini teknik deteksi dan lokalisasi bunyi yang digunakan adalah SELDnet. Selanjutnya kelompok ketiga ini akan disebut sebagai *SELDnet only*.

Tabel IV.5 Hasil Perbandingan Performa Pemisahan SELDnet dengan SELDnet only

	Class aware localization scores		Location-aware detection scores		seld_score
	DOA_error	Frame_recall	F_Score	Error Rate	
Separation SELDnet combination dataset	15,5	46,3	46,3	0,72	0,43
Separation SELDnet full augmantion dataset	12,5	81,5	71,1	0,41	0,24
SELDnet only	25,1	56,9	32,6	0,76	0,5

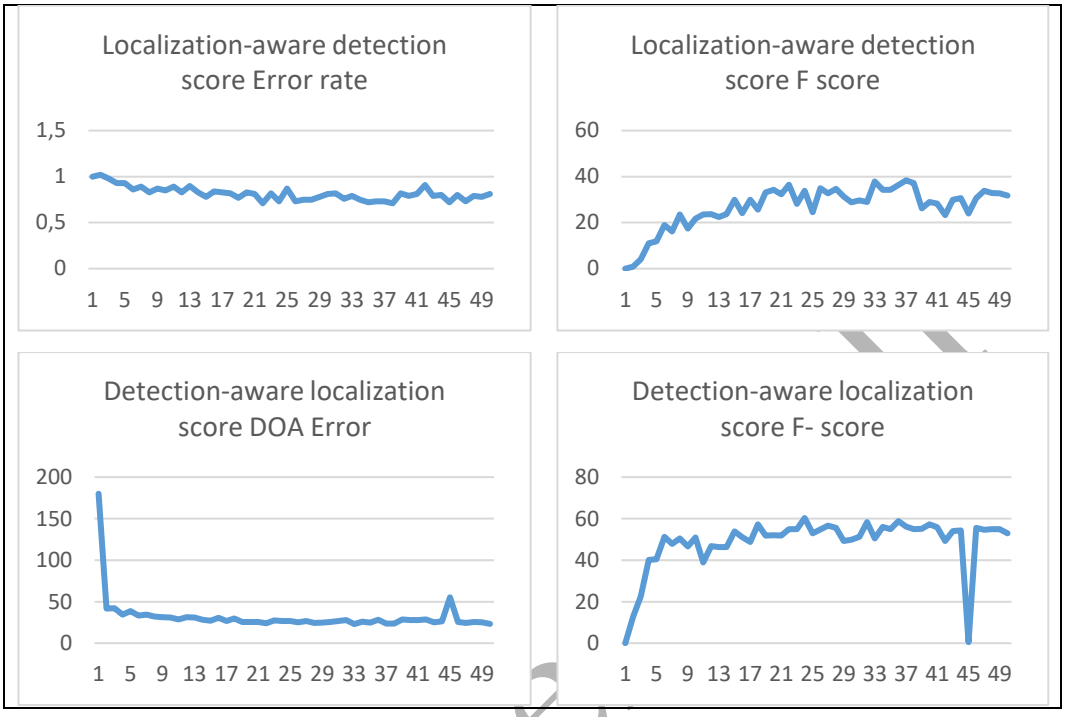


Gambar IV.12 Grafik Hasil Performa Pemisahan SELDnet dengan SELDnet only

Ketiga kelompok data latih di atas akan diukur performa dan akurasinya menggunakan teknik pengukuran *class-aware localization scores* dan *localitatio-aware detection score*, juga diukur *seld-score*-nya. Hasil pengujian dapat dilihat pada Tabel IV.5 dan Gambar IV.12. Berdasarkan Tabel IV.5 dapat dilihat bahwa tingkat *error* untuk teknik Pemisahan SELDnet lebih rendah dibandingkan tingkat *error* untuk teknik hanya SELDnet saja. Begitu juga untuk hasil *frame recall*-nya teknik Pemisahan SELDnet juga memberikan hasil yang lebih baik dibanding teknik hanya SELDnet. Hasil untuk lokalisasi bunyi dengan memperhatikan skor deteksinya, yaitu nilai *F-score*, serta nilai *error rate* dari Pemisahan SELDnet dengan data augmentasi memiliki nilai terbaik, begitu juga untuk skor SELD-nya memiliki nilai terkecil. Keseluruhan perbandingan performa ini menunjukkan bahwa teknik Pemisahan SELDnet memiliki akurasi terbaik dengan data augmentasi yang telah berhasil meningkatkan juga jumlah variasi dari data bunyi tumpang tindih. Berdasarkan eksperimen yang telah dilakukan, teknik Pemisahan SELDnet menunjukan telah berhasil meningkatkan akurasi untuk deteksi dan lokalisasi bunyi tumpang tindih tanpa mempengaruhi tingkat akurasi untuk bunyi tunggal. Hal ini juga menunjukkan bahwa teknik Pemisahan SELDnet bersifat adapatif terhadap data bunyi tumpang tindih maupun data bunyi tunggal.

Berdasarkan hasil perbandingan, bisa dilihat hasil dari Pemisahan SELDnet menunjukan peningkatan performa, untuk itu bisa ditunjukan juga hasil yang lebih

detail dari eksperimen Pemisahan SELDnet. Berdasarkan penjabaran eksperimen yang lebih detail bisa dilihat parameter yang mempengaruhi hasil eksperimen sehingga dapat dianalisa hal-hal yang mempengaruhi peningkatan performa.

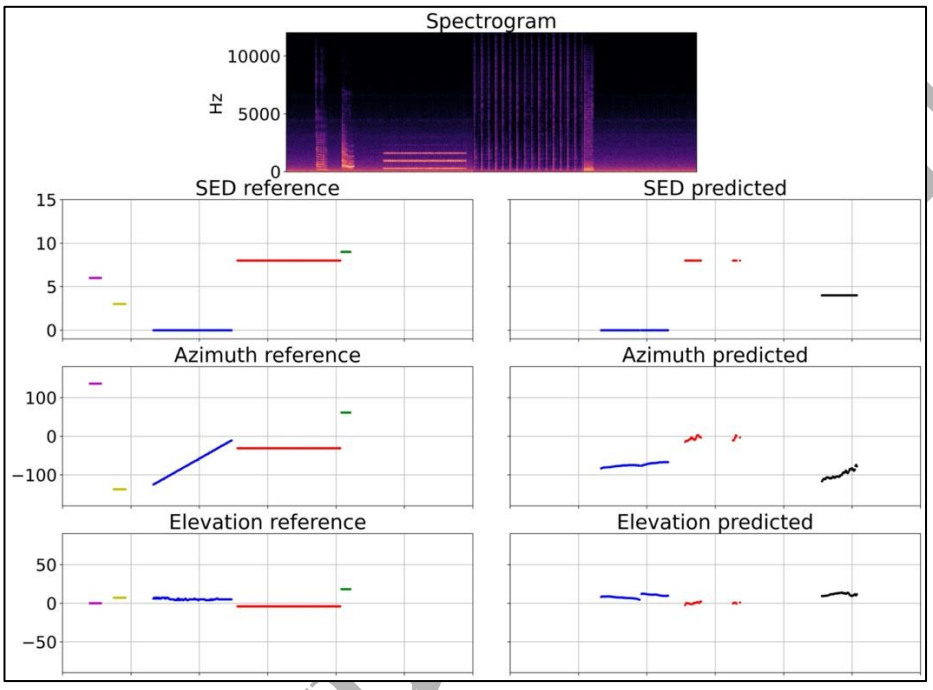


Gambar IV.13 Detail eksperimen Pemisahan SELDnet dengan data augmentasi TAU NIGENS 2020

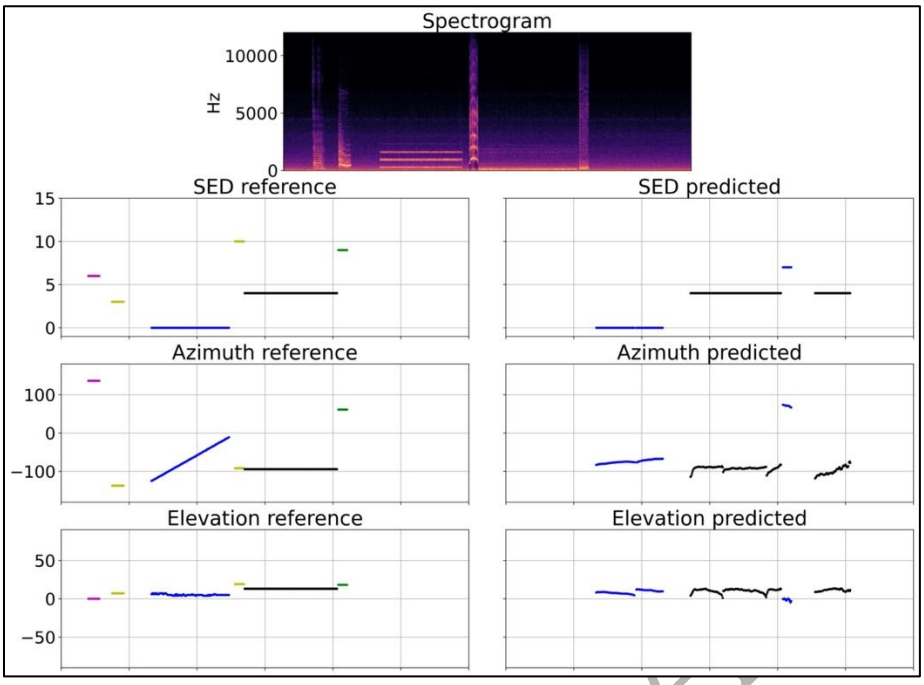
Detail dari eksperimen kelompok pertama yaitu Pemisahan SELDnet dengan data augmentasi TAU NIGENS 2020 dapat dilihat pada Gambar IV.13. Gambar IV.13 menunjukan perubahan dari skor deteksi pada *localization-aware* dan skor lokalisasi pada *detection-aware* untuk setiap epochnya. Pada tingkat *error* memiliki kecenderungan menurun untuk kedua jenis skor, hal ini menandakan kesalah deteksi dan lokalisasi lebih sedikit terjadi selama tahap pelatihan model *neural network*-nya. Begitu juga pada nilai *F score* dikedua jenis skornya, pada setiap epoch mengalami peningkatan, walaupun di epoch terakhirnya terjadi penurunan secara singkat.

Hasil akurasi dari Pemisahan SELDnet juga digambarkan pada grafik akurasi perbandingan antara hasil prediksi dengan acuannya. Hasil ini dapat dilihat pada Gambar IV.14 dan Gambar IV.15 yang menunjukan performa dari Pemisahan SELDnet memiliki akurasi yang meningkat signifikan jika dibandingkan SELDnet

only dan dengan data asli TAU NIGENTS 2020. Hal ini dapat dilihat karena bentuk grafik antara *reference* dan *predicted* yang cenderung sama atau mirip. Pada gambar sumbu x adalah index waktu dan sumbu y adalah labelnya. Sedangkan warna garis menunjukan kelompok jenis bunyinya. Berdasarkan kedua gambar dapat dilihat kemiripan gambarnya menunjukan hasil pemisahan telah dapat digunakan untuk deteksi dan lokalisasi bunyi.

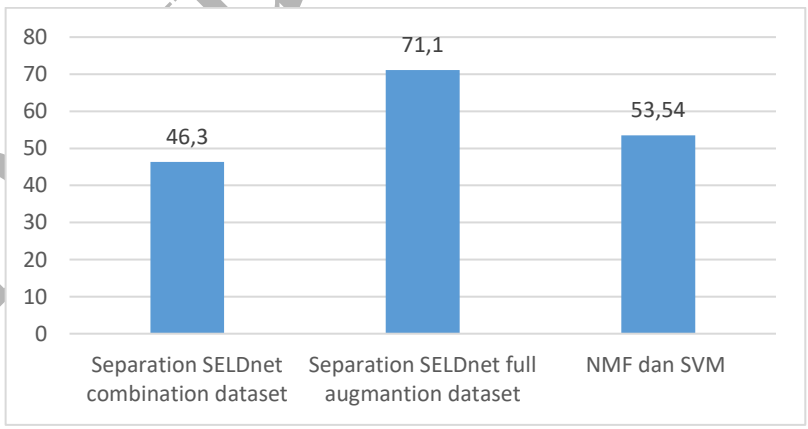


Gambar IV.14 Grafik *reference* dan *predicted* hasil eksperimen Pemisahan SELDnet



Gambar IV.15 Grafik *reference* dan *predicted* hasil eksperimen Pemisahan SELDnet

Sebagai tambahan analisa, hasil eksperimen B dan D juga dibandingkan, hasilnya dapat dilihat Gambar IV.16. Berdasarkan gambar dapat dilihat bahwa teknik Pemisahan SELDnet dengan *full augmentation dataset* memiliki akurasi terbaik. Namun jika menggunakan *combination dataset* hasilnya masih lebih rendah dibanding NMF dan SVM.



Gambar IV.16 Hasil perbandingan akurasi

Bab V Kesimpulan dan Saran

V.1 Kesimpulan

Berdasarkan hasil eksperimen dapat disimpulkan beberapa hal, antara lain:

1. Penambahan teknik pemisahan bunyi menggunakan *Nonnegative Matrix Factorization* pada sistem deteksi dan lokalisasai bunyi SELDnet mampu meningkatkan performa sistem pengenalan bunyi, terutama pada bunyi tumpang tindihnya. Hal ini dapat dilihat dari kenaikan performa rata-rata sebesar 26 persen untuk *F-score* dan menurunnya nilai *error* sebesar 0,2 persen.
2. Proses augmentasi data yang memperbanyak jumlah varian dari bunyi tumpang tindih juga mempengaruhi secara positif performa dari pemisahan bunyi tumpang tindih *Nonnegative Matrix Factorization* serta pengenalan bunyi SELDnet. Hal ini dapat dilihat dari peningkatan hasil akurasi antara data keseluruhan augmentasi dengan data aslinya yaitu sebesar 39 persen peningkatan nilai *F-score*.
3. Berdasarkan pengujian, juga dapat disimpulkan bahwa teknik pemisahan bunyi menggunakan *Nonnegative Matrix Factorization* dengan pengenalan bunyi SELDnet lebih bersifat adaptif untuk kondisi data yang tunggal maupun tumpang tindih, sehingga peluang untuk diterapkan pada kondisi nyata lebih besar.

V.2 Saran

Berdasarkan eksperimen masih terdapat peluang penelitian lainnya. Teknik pemisahan bunyi dan bentuk topologi *Neural Network* lainnya yang diterapkan pada sistem pengenalan bunyi lainnnya dapat dikembangkan dan menjadi peluang penelitian ke depannya. Penggunaan data dengan berbagai kondisi tumpang tindih juga dapat digunakan pada sistem yang dikembangkan pada penelitian ini. Pengujian dengan proses data augmentasi yang lebih kompleks juga dapat dilakukan untuk meningkatkan akurasi, misalnya dengan menambahkan kondisi bunyi dengan gema yang lebih bervariasi. Penerapan sistem pada kondisi nyata juga dapat diujikan agar dapat melihat kembali kekurangan-kekurangan dari sistem yang dibangun. Pengembangan kearah sistem aplikatif juga dapat dimulai berdasarkan

hasil penelitian yang dilakukan ini. Pengukuran dengan memperhatikan *computing time* juga bisa ditambahkan sebagai alat ukur kinerja sistem.

DAFTAR PUSTAKA

Adavanne, S., Pertila, P., and Virtanen, T. (2017): Sound event detection using spatial features and convolutional recurrent neural network, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 771–775. https://doi.org/10.1109/ICASSP.2017.7952260

Adavanne, S., Politis, A., Nikunen, J., and Virtanen, T. (2019): Sound Event Localization and Detection of Overlapping Sources Using Convolutional, *IEEE Journal of Selected Topics in Signal Processing*, **13**(1), 34–48. https://doi.org/10.1109/JSTSP.2018.2885636

Adavanne, S., Politis, A., and Virtanen, T. (2017): Direction of arrival estimation for multiple sound sources using convolutional recurrent neural network, *CoRR*, retrieved from internet: http://arxiv.org/abs/1710.10059, **abs/1710.1**.

Adavanne, S., Politis, A., and Virtanen, T. (2018): Multichannel Sound Event Detection Using 3D Convolutional Neural Networks for Learning Inter-channel Features, *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–7. https://doi.org/10.1109/IJCNN.2018.8489542

Avenue, M., and Hill, M. (n.d.): A STUDY OF SPEECH RECOGNITION FOR CHILDREN AND THE ELDERLY . J a y G . Wilpon , Claw N . Jacobsen * AT & T Bell Laboratories, 5–8.

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., and K. Shikano (2004): Acoustic models of the elderly for large-vocabulary continuous speech recognition, *Electronics and Communications in Japan*, **87**, 49–57.

Berg, R. E., Stork, D. G., and Holmes, B. (1982): The Physics of Sound, , retrieved September 9, 2017, from internet: http://aapt.scitation.org/doi/10.1119/1.12960. https://doi.org/10.1119/1.12960

Bisot, V., Essid, S., and Richard, G. (2017): Overlapping sound event detection with supervised Nonnegative Matrix Factorization, *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 31–35. https://doi.org/10.1109/ICASSP.2017.7951792

Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H., and Virtanen, T. (2017): Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**(6), 1291–1303. https://doi.org/10.1109/TASLP.2017.2690575

Cao, Y., Kong, Q., Iqbal, T., An, F., Wang, W., and Plumbley, M. D. (2019): Polyphonic Sound Event Detection and Localization Using a Two-Stage Strategy, *Detection and Classification of Acoustic Scenes and Events 2019*, 30–34.

Chakrabarty, S., and Habets, E. A. P. (2017): Multi-Speaker Localization Using Convolutional Neural Network Trained with Noise, *Proc. Neural Inf. Process. Syst*, retrieved from internet: http://arxiv.org/abs/1712.04276, **abs/1712.0**.

Chen, T.-E., Yang, S.-I., Ho, L.-T., Tsai, K.-H., Chen, Y.-H., Chang, Y.-F., Lai, Y.-H., Wang, S.-S., Tsao, Y., and Wu, C.-C. (2017): S1 and S2 Heart Sound

Recognition Using Deep Neural Networks, *IEEE Transactions on Biomedical Engineering*, **64**(2), 372–380. https://doi.org/10.1109/TBME.2016.2559800

Cheong Took, C., Sanei, S., Rickard, S., Chambers, J., and Dunne, S. (2008): Fractional delay estimation for blind source separation and localization of temporomandibular joint sounds, *IEEE Transactions on Biomedical Engineering*, **55**(3), 949–956. https://doi.org/10.1109/TBME.2007.909534

Chia Ai, O., Hariharan, M., Yaacob, S., and Sin Chee, L. (2012): Classification of speech dysfluencies with MFCC and LPCC features, *Expert Systems with Applications*, **39**(2), 2157–2165. https://doi.org/10.1016/j.eswa.2011.07.065

Chung, J., Gulcehre, C., Cho, K., and Yoshua Bengio (2014): Gated Recurrent Neural Networks on Sequence Modeling, 1–9.

Darji, M. (2017): Audio Signal Processing : A Review of Audio Signal Classification Features, (May).

Dibiase, J. H. (August 2000): *A high-accuracy, low-latency technique for talker localization in reverberant environments using microphone arrays*, BROWN UNIVERSITY.

Ferguson, E. L., Williams, S. B., and Jin, C. T. (2018): Sound Source Localization in a Multipath Environment Using Convolutional Neural Networks, *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process*, retrieved from internet: http://arxiv.org/abs/1710.10948, 2386–2390.

Ganchev, T. D. (2005): *Speaker Recognition, Wire Communications Laboratory Department of Computer and Electrical Engineering University of Patras Greece*.

Gao, B., Woo, W. L., and Dlay, S. S. (2011): Adaptive sparsity non-negative matrix factorization for single-channel source separation, *IEEE Journal on Selected Topics in Signal Processing*, **5**(5), 989–1001. https://doi.org/10.1109/JSTSP.2011.2160840

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., Ritter, M., View, M., and York, N. (2017): Audio Set : an Ontology and Human-Labeled Dataset for Audio Events, 776–780.

Gilke, M., Kachare, P., Kothalikar, R., and Rodrigues, V. P. (2012): MFCC-based Vocal Emotion Recognition Using ANN, *2012 International Conference on Electronics Engineering and Informatics (ICEEI 2012)*, **49**(Iceei), 150–154. https://doi.org/10.7763/IPCSIT.2012.V49.27

Gunn, S. R. (1998): Support Vector Machines for Classification and Regression.

Han, T. J., Kim, K. J., and Park, H. (2015): Location estimation of predominant sound source with embedded source separation in amplitude-panned stereo signal, *IEEE Signal Processing Letters*, **22**(10), 1685–1688. https://doi.org/10.1109/LSP.2015.2424991

HAPILABS (2016): Hapifork.

Harma, A., McKinney, M. F., and Skowronek, J. (2005): Automatic surveillance of the acoustic activity in our living environment, *2005 IEEE International*

Conference on Multimedia and Expo, 4 pp. https://doi.org/10.1109/ICME.2005.1521503

He, W., Motlicek, P., and Odobez, J.-M. (2017): Deep Neural Networks for Multiple Speaker Detection and Localization, *Proc. Int. Conf. Robot. Autom*, retrieved from internet: http://arxiv.org/abs/1711.11565, **abs/1711.1**, 74–79.

Hirvonen, T. (2015): Classification of Spatial Audio Location and Content Using Convolutional Neural Networks, *Proc. Audio Eng. Soc.*

Huang, Y., Benesty, J., Elko, G. W., and Mersereati, R. M. (2001): Real-time passive source localization: a practical linear-correction least-squares approach, *IEEE Trans. Speech Audio Process.*, **9**, 943–956.

Istrate, D., Vacher, M., and Serignat, J. F. (2006): Generic implementation of a distress sound extraction system for elder care, *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, 3309–3312. https://doi.org/10.1109/IEMBS.2006.259469

James (2010): Mixing two audio files together with python, retrieved September 9, 2018, from internet: https://stackoverflow.com/questions/4039158/mixing-two-audio-files-together-with-python.

Khan, M. S., Naqvi, S. M., and Chambers, J. A. (2013): A new cascaded spectral subtraction approach for binaural speech dereverberation and its application in source separation, *EEE International Conference on Acoustics Speech and Signal Processing*.

Kim, H. G., Moreau, N., and Sikora, T. (2004): Audio classification based on MPEG-7 spectral basis representations, *IEEE Transactions on Circuits and Systems for Video Technology*, **14**(5), 716–725. https://doi.org/10.1109/TCSVT.2004.826766

Kwon, B., Park, Y., and Park, Y. S. (2008): Sound source localization for robot auditory system using the summed GCC method, *2008 International Conference on Control, Automation and Systems, ICCAS 2008*, **1**(1), 241–245. https://doi.org/10.1109/ICCAS.2008.4694557

Lee, D. D., Hill, M., and Seung, H. S. (2001): Algorithms for Non-negative Matrix Factorization, *Advances in neural information processing systems*, 556–562.

Li, Y., Woodruff, J., and Wang, D. (2009): Monaural musical sound separation based on pitch and common amplitude modulation, *IEEE Transactions on Audio, Speech and Language Processing*, **17**(7), 1361–1371. https://doi.org/10.1109/TASL.2009.2020886

Lim, H., Park, J., Lee, K., and Han, Y. (2017): RARE SOUND EVENT DETECTION USING 1D CONVOLUTIONAL RECURRENT NEURAL NETWORKS Cochlear . ai , Seoul , Korea Music and Audio Research Group , Seoul National University , Seoul , Korea, (November), 2–6.

Liu, H., Wu, Z., Li, X., Cai, D., and Huang, T. S. (2012): Constrained Nonnegative Matrix Factorization for Image Representation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(7), 1299–1311. https://doi.org/10.1109/TPAMI.2011.217

Luo, Y., and Mesgarani, N. (2019): Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation, *IEEE/ACM Transactions on Audio Speech and Language Processing*, **27**(8), 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167

Manilow, E., Seetharaman, P., and Pardo, B. (2018a): The Northwestern University Source Separation Library., *Proceedings of the 19th International Society of Music Information Retrieval Conference*, 297–305.

Manilow, E., Seetharaman, P., and Pardo, B. (2018b): The Northwestern University Source Separation Library, *Proceedings of the 19th International Society of Music Information Retrieval Conference*, Paris.

Maunder, D., Ambikairajah, E., Epps, J., and Celler, B. (2008): Dual-microphone Sounds of Daily Life classification for telemonitoring in a noisy environment., *Conference Proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, **2008**(1), 4636–9. https://doi.org/10.1109/IEMBS.2008.4650246

Mesaros, A., Adavanne, S., Politis, A., Heittola, T., and Virtanen, T. (2019): Joint Measurement of Localization and Detection of Sound Events, *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, 333–337.

Mesaros, A., Diment, A., Elizalde, B., Heittola, T., Vincent, E., Raj, B., and Virtanen, T. (2019): Sound Event Detection in the DCASE 2017 Challenge, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **27**(6), 992–1006. https://doi.org/10.1109/TASLP.2019.2907016

Mogi, R., and Kasai, H. (2012): Noise-Robust environmental sound classification method based on combination of ICA and MP features, *Artificial Intelligence Research*, **2**(1), 107–121. https://doi.org/10.5430/air.v2n1p107

Mondal, A., Banerjee, P., and Somkuwar, A. (2017): Enhancement of lung sounds based on empirical mode decomposition and Fourier transform algorithm, *Computer Methods and Programs in Biomedicine*, **139**, 119–136. https://doi.org/10.1016/J.CMPB.2016.10.025

Mueller, M. (2015): *Fundamentals of Music Processing*.

Niranjani, K., and Vani, K. (2017): Unsupervised linear spectral unmixing of multispectral images using the NMF and modified-multilayer NMF algorithms, *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*, 1440–1444. https://doi.org/10.1109/ICPCSI.2017.8391950

Pain, H. J. (2005): *The Physics of Vibrations and Waves* (6th ed.), John Wiley & Sons Ltd, 570. https://doi.org/10.1002/0470016957

Park, D. S., Chan, W., Zhang, Y., Chiu, C., Zoph, B., Cubuk, E. D., and Le, Q. V (2020): SpecAugment : A Simple Data Augmentation Method for Automatic Speech SpecAugment : A Simple Data Augmentation Method for Automatic Speech Recognition, (April 2019).

Piczak, K. J. (2014): ESC-50: Dataset for Environmental Sound Classification.

Piczak, K. J. (2015): ESC: Dataset for Environmental Sound Classification, *Proceedings of the 23rd ACM International Conference on Multimedia, MM 2015*, 1015–1018. https://doi.org/10.1145/2733373.2806390

Politis, A., Adavanne, S., and Virtanen, T. (2020): A Dataset of Reverberant Spatial Sound Scenes with Moving Sources for Sound Event Localization and Detection, retrieved from internet: http://arxiv.org/abs/2006.01919, (November), 165–169.

Rajabi, R., and Ghassemian, H. (2014): Multilayer structured NMF for spectral unmixing of hyperspectral images, *2014 6th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 1–4. https://doi.org/10.1109/WHISPERS.2014.8077521

Robert, J. (2011): Pydub, , retrieved August 10, 2018, from internet: https://github.com/jiaaro/pydub#installation.

Roy, R., and Kailath, T. (1989): ESPRIT-estimation of signal parameters via rotational invariance techniques, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **37**(7), 984–995. https://doi.org/10.1109/29.32276

Russell, S., and Norvig, P. (2016): *Arficial Intelligence A Modern Approach* (Third) (S. Russel and P. Norvig, Eds.), Person, London, 727–728.

Sailor, H. B., Agrawal, D. M., and Patil, H. A. (2017): Unsupervised filterbank learning using Convolutional Restricted Boltzmann Machine for environmental sound classification, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, **2017-Augus**(1), 3107–3111. https://doi.org/10.21437/Interspeech.2017-831

Salman Khan, M., Yu, M., Feng, P., Wang, L., and Chambers, J. (2015): An unsupervised acoustic fall detection system using source separation for sound interference suppression, *Signal Processing*, **110**, 199–210. https://doi.org/10.1016/j.sigpro.2014.08.021

Scenes, A. (2017): A REPORT ON SOUND EVENT DETECTION WITH DIFFERENT BINAURAL FEATURES Sharath Adavanne , Tuomas Virtanen Department of Signal Processing , Tampere University of Technology, (November), 14–17.

Schmidt, R. (1986): Multiple emitter location and signal parameter estimation, *IEEE Transactions on Antennas and Propagation*, **34**(3), 276–280. https://doi.org/10.1109/TAP.1986.1143830

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001): Estimating the Support of a High-Dimensional Distribution, *Neural Computation*, **13**(7), 1443–1471. https://doi.org/10.1162/089976601750264965

Scikit-learn developer (n.d.): Clustering, retrieved October 23, 2019, from internet: https://scikit-learn.org/stable/modules/clustering.html.

Shah, G., Member, S., Koch, P., and Papadias, C. B. (2015): On the Blind Recovery of Cardiac and Respiratory Sounds, **19**(1), 151–157.

Shimada, K., Takahashi, N., Takahashi, S., and Mitsufuji, Y. (2020): *Sound Event Localization and Detection Using Activity-Coupled Cartesian Doa Vector and Rd3net*, 1–4.

Teutsch, H. (2007): *Modal Array Signal Processing: Principles and Applications of Acoustic Wavefield Decompositio* (1st ed.), Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-540-40896-3

Tian, D., Xu, X., and Wang, X. (2017): An Improved Activity Recognition Method Base G on Smart Watch Data, *2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 756. https://doi.org/10.1109/CSE-EUC.2017.148

Tripathi, A. M., and Baruah, R. D. (2017): Acoustic event classification using Cauchy Non-negative matrix factorization and fuzzy rule-based classifier, *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–6. https://doi.org/10.1109/FUZZ-IEEE.2017.8015584

Vacher, M., Fleury, A., Portet, F., Serignat, J.-F., and Noury, N. (2010): Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living, *New Developments in Biomedical Engineering*, 645–673. https://doi.org/10.5772/7596

Valin, J. M. (2007): On Adjusting the Learning Rate in Frequency Domain Echo Cancellation With Double-Talk, *IEEE Transactions on Audio, Speech, and Language Processing*, **15**(3), 1030–1034. https://doi.org/10.1109/TASL.2006.885935

Vesperini, F., Vecchiotti, P., Principi, E., Squartini, S., and Piazza, F. (2016): A neural network based algorithm for speaker localization in a multi-room environment, *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1–6. https://doi.org/10.1109/MLSP.2016.7738817

Yang, L.-C., and Lerch, A. (2020): Remixing Music with Visual Conditioning, *2020 IEEE International Symposium on Multimedia (ISM)*, 181–188. https://doi.org/10.1109/ISM.2020.00039

Yilmaz, Ö., and Rickard, S. (2004): Blind separation of speech mixtures via time-frequency masking, *IEEE Transactions on Signal Processing*, **52**(7), 1830–1846. https://doi.org/10.1109/TSP.2004.828896

Zafeiriou, S., Tefas, A., Buciu, I., and Pitas, I. (2006): Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification, *IEEE Transactions on Neural Networks*, **17**(3), 683–695. https://doi.org/10.1109/TNN.2006.873291

Zhang, J., Ding, W., and He, L. (2019): *Data Augmentation and Prior Knowledge-Based Regularization for Sound Event Localization and Detection*, 1–5.

LAMPIRAN

Lampiran A Tabel Kelas Bunyi pada NIGENS

No	Kelas	Deskripsi	Total file	Total durasi	Rata-rata durasi
1	Alaram (*Alarm)*	Berbagai jenis bunyi dari alaram kebakaran model lama hingga yang berupa alaram elektronik. Sebagian besar bernada tinggi, terpotong-potong, kontinu, terstruktur, meraung-raung secara terus menerus.	49	15mnt:48dtk	19,4dtk
2	Tangisan bayi (*Baby crying*)	Tangisan bayi. Sebagian besar berupa rentetan tangisan, bisa juga bunyi isak tunggal dan memekik.	40	18mnt:02dtk	27,1dtk
3	Tabrakan (*Crash*)	Rangkaian kecelakaan, tabrakan keras, mirip bunyi *noise*, tapi kejadian mendadak, pecah, bunyi tunggal. Energi bunyi tinggi di antara frekuensi dengan skala yang lebar.	50	8mnt:10dtk	9,8dtk
4	Gonggongan anjing (*Dog barking*)	Bunyi anjing menggonggong, sebagian besar beberapa kali dalam satu waktu. Puncak energi bunyi sekitar 1kHz, pendek,	45	8mnt:43dtk	11,6dtk

No	Kelas	Deskripsi	Total file	Total durasi	Rata-rata durasi
		kejadian terpotong-potong			
5	Mesin (*Engine*)	Berbagai jenis bunyi mesin menyala dengan durasi panjang dan terus menerus, kondisi mesin sedang tidak aktif atau kecepatan bunyi mesin sedang berubah.	39	34mnt:40dtk	53,6dtk
6	Teriakan wanita (*females scream*)	Teriakan wanita yang singkat pendek dan tungga, bernada tinggi, titik puncak energi bunyinya sekitar 1.8kHz.	45	2mnt:46dtk	3,7 dtk
7	Suara wanita berbicara	Suara wanita berbicara dengan lembut mengucapkan kalimat yang pendek.	100	4m:53dtk	2,9dtk
8	Api (*fire*)	Bunyi api terbakar yang panjang dan terus menerus. Bunyinya mirip seperi bunyi *noise*, tapi dengan energi bunyi yang tinggi pada frekuensi rendah.	51	45mnt:20dtk	53,4dtk
9	Langkah kaki (*footsteps*)	Berbagai jenis bunyi langkah kaki manusia berjalan pada beberapa jenis permukaan, seperti	42	18mnt:43dtk	26,8dtk

No	Kelas	Deskripsi	Total file	Total durasi	Rata-rata durasi
		kayu hingga salju. Berupa rangkaian kejadian yang pendek.			
10	Umum (*general*)	Bunyi-bunyi diluar kelas bunyi yang ada. Terpotong-potong dan kontinu, kejadian tunggal maupun sekuensial, memuncak atau melebar.	303	1jam:31mnt:49dtk	18,2dtk
11	Ketukan (*knocking*)	Bunyi ketukan pada sesuatu, sebagian besar berupa pintu. Berupa bunyi sekuen dari kejadian yang sangat singkat. Sebagian besar energi bunyinya pada *bandwith* yang rendah	40	1mnt:42dtk	2,6dtk
12	Teriakan pria (*male scream*)	Terikan yang pendek dan tunggal dari seorang pria. Puncak energi bunyinya sekitar 1,2kHz	31	3mnt:16dtk	6,4dtk
13	Suara pria berbicara (*male speech*)	Suara pria berbicara kalima pendek secara pelan.	100	4mnt:15dtk	2,6dtk
14	Telepon berdering (*phone ringing)*	Sebagian besar bunyi berdering dari telepon klasik. Bunyi sekuen dari dering yang panjang.	40	12mnt:16dtk	18,4dtk

No	Kelas	Deskripsi	Total file	Total durasi	Rata-rata durasi
15	Piano	Bunyi piano dimainkan. Bunyinya jenis sekuen *polyphonic* dan *monophonic*.	42	14mnt:33dtk	20,8dtk

Lampiran B Matriks Penelitian Acuan

Topik Pemisahan Bunyi				
No	Judul, Penulis, Tahun	Permasalahan	Teknik	Hasil
1.	*Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation,* Yi Luo, Nima Mesgarani, 2019 (Luo dan Mesgarani, 2019).	Memisahkan suara orang berbicara yang saling tumpang tindih.	Mengembangkan metode pemisahan Conv-TasNet-cLN.	Hasilnya terjadi peningkatan akurasi jika dibandingkan dengan teknik STFT. Namun dari segi kecepatan dan penerapan dengan kondisi nyata masih bisa dikembangkan.
2	*The Northwestern University Source Separation Library*, Ethan Manilow, Prem Seetharaman, Bryan Pardo, 2018 (Manilow dkk., 2018)	Melakukan publikasi *library nussl* yang berisi teknik pemisahan bunyi dan pengolahan dasar bunyi lainnya.	Terdapat beberapa teknik antara lain: *Repetition, Matrix Decomposition, Other Force/Background, Component Analysis, Spatialization, Benchmarking*	Proses evaluasinya yang dilakukan ada nilai *precision, recall, F-Score, accuracy*, dan menunjukkan *library nussl* sudah dapat digunakan untuk penelitian sejenisnya. Pengujian menggunakan data bunyi musik

Topik Pemisahan Bunyi				
No	Judul, Penulis, Tahun	Permasalahan	Teknik	Hasil
3	*Remixing Music with Visual Conditioning*, Li-Chia Yang dan Alexander Lerch, 2020 (L. Yang dan Lerch, 2020)	Melakukan *remixing music* dari berdasarkan bunyi instrumen dan gambar instrumennya dalam suatu citra digital dan video.	Teknik pemisahan bunyinya adalah Deep U-Net dan teknik analisa videonya adalah Dilated ResNet-18	Pengujian dilakukan menggunakan dataset MUSIC yang kemudian diukur hasil *mixing music*-nya. Diukur menggunakan nilai SNR dari hasil *mixing*. Hasilnya menunjukkan bahwa model yang bentuk dapat mengontrol nilai volume dari hasil *mixing.*
4	*Blind Separation of Speech Mixtures via Time-Frequency Masking*, O¨zgu¨r Yılmaz dan Scott Rickard, 2004 (Yilmaz and Rickard, 2004)	Memisahkan suara orang berbicara yang sumber suaranya lebih dari satu dalam waktu yang sama.	Teknik pemisahan yang diguanakan adalah *approximate W-disjoint orthogonality*.	Teknik bekerja dengan baik pada data sintetis, tapi perlu dikembangkan agar lebih fleksibel terhadap kondisi frekuensi bunyinya.
5	*Separation of Moving Sound Sources Using Multichannel NMF and*	Memisahkan bunyi yang berasal dari sumber bunyi yang bergerak. Jenis bunyi yang dipisahkan	Teknik pemisahan yang digunakan adalah NMF yang	Teknik diukur menggunakan SDR, SIR, SAR dan ISR. Nilai *error rate*-nya adalah 8.8 derajat dan *recall*

Topik Pemisahan Bunyi				
No	Judul, Penulis, Tahun	Permasalahan	Teknik	Hasil
	Acoustic Tracking, Joonas Nikunen, Aleksandr Diment, dan Tuomas Virtanen, 2017	adalah bunyi orang berbicara sebanyak empat sumber bunyi yang berbeda bergerak bersamaan menuju *microphone.*	dimodifikasi pada jumlah *channel*-nya.	*rate*-nya 79% yang mana ini nilainya mirip dari teknik pembandingnya.
Topik Deteksi, Lokalisasi, dan Klasifikasi Bunyi				
No	Judul, Penulis, Tahun	Permasalahan	Teknik	Hasil
6	*ESC: Dataset for Environmental Sound Classification*, Karol J. Piczak, 2015 (Piczak, 2015)	Melakukan publikasi data bunyi lingkungan yang diimplementasikan pada kasus klasifikasi bunyi.	Teknik klasifikasi yang digunakan adalah *Random fores, SVM, dan k-NN.*	Hasil untuk random forest (44.3%), SVM (39.6%) dan k-NN (32.2%).
7	*Sound Event Localization and Detection of Overlapping Sources Using Convolutional Recurrent Neural Networks*, Adavanne, 2017	Melakukan lokalisasi dan deteksi bunyi dengan berbagai kondisi bunyi: tunggal, tumpang tindih dan bergema.	Teknik yang digunakan adalah *Sound Event Localization and Detection net.* Teknik ini menggunakan	Hasil akurasi pada data tunggal lebih baik dibanding data bunyi tumpang tindih, selisih akurasinya di atas 20 persen.

Topik Pemisahan Bunyi				
No	Judul, Penulis, Tahun	Permasalahan	Teknik	Hasil
	(Adavanne, Pertila, dkk., 2017).		beberapa teknik jaringan syaraf tiruan seperti, RNN dan FNN.	
8	*Sound Source Localization for Robot Auditory System Using the Summed GCC Method*, Byoungho Kwon, 2008 (Kwon dkk., 2008)	Mengimplementasikan sistem pengenalan lokasi sumber bunyi pada robot.	Teknik yang digunakan adalah *The Summed* GCC.	Dengan tambahan alat *microphone* 3D mampu melakukan lokalisasi bunyi dengan cukup akurat.
9	*Polyphonic Sound Event Detection and Localization using a Two-Stage Strategy*, Yin Cao, dkk, 2019(Cao dkk., 2019)	Mengembangkan metode pengenalan jenis bunyi dan mengestimasi lokasi spasial dan temporalnya pada bunyi tumpang tindh.	Teknik yang digunakan adalah *two-stage sound event detection and localization network.*	Menunjukan akurasi yang cukup baik untuk DOA.

Topik Pemisahan Bunyi				
No	Judul, Penulis, Tahun	Permasalahan	Teknik	Hasil
10	*Noise-Robust environmental sound classification method based on combination of ICA and MP features*, Reona Mogi dan Hiroyuki Kasai, 2013 (Mogi dan Kasai, 2012).	Permasalahan yang diangkat adalah klasifikasi bunyi lingkungan dengan *noise*.	Teknik yang digunakan adalah *Independent Component Analysis* (ICA) dan *Matching Pursuit* (MP).	Hasilnya akurasi meningkat sebesar 8% jika dibandingkan dengan teknik MFCC.

Lampiran C Hasil akurasi klasifikasi data tumpang tindih (dalam persen (%))

Jenis bunyi	can_opening	clock_alarm	clock_tick	door_wood_creaks	door_wood_kncok	glass_breaking	keyboard_typing	mouse_click	vacuum_cleaner	washing_machine	Rata-rata
breathing	86.67	**71.43**	90.79	**77.78**	83.33	89.74	92.50	87.14	**71.25**	88.16	83.88
brushing	80.88	**68.92**	**76.32**	93.42	83.33	80.77	80.00	**79.73**	**78.75**	**78.95**	80.11
clapping	**79.17**	94.74	**56.58**	82.90	**70.00**	**71.25**	88.00	**65.28**	86.25	92.31	**78.65**
coughing	100.00	82.35	82.90	100.00	97.83	87.84	91.67	**65.00**	**72.50**	**70.51**	85.06
crying_baby	91.43	**75.68**	**76.32**	86.84	82.90	91.25	96.15	**65.28**	**68.75**	89.74	82.43
drinking	81.48	**68.57**	97.30	81.08	83.33	98.65	84.09	**75.71**	**58.75**	**67.95**	**79.69**
footstep	88.24	**75.00**	98.68	88.46	81.94	82.50	93.75	**79.17**	**71.25**	**76.32**	83.53
laughing	90.32	87.14	**65.79**	100.00	**75.86**	**79.49**	97.50	**66.67**	**70.00**	94.87	82.76
sneezing	100.00	**69.36**	100.00	81.43	100.00	91.67	**75.00**	**73.33**	**61.25**	**63.16**	81.52
snoring	91.67	80.26	96.05	96.15	95.95	85.00	96.30	**78.38**	83.75	98.72	90.22
Rata-rata	88.99	**77.34**	84.07	88.81	85.45	85.82	89.50	**73.57**	**72.25**	82.07	

Lampiran D Daftar Publikasi

Ranny, Lestari, D. P., dan Mengko, T. L. E. R. (2024): Separation Sound Event Localization and Detection using Neural Network and Time Frequency Masking, *International Journal of Innovative Computing, Information and Control*, **20**.

Ranny, R., Lestari, D. P., Latifah Erawati Rajab, T., dan Suwardi, I. S. (2019): Separation of Overlapping Sound using Nonnegative Matrix Factorization, *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 424–429. https://doi.org/10.1109/ISRITI48646.2019.9034580

Ranny, R., Suwardi, I. S., Rajab, T. L. E., dan Lestari, D. P. (2019): Kajian Penelitian Pemrosesan Bunyi dan Aplikasinya pada Teknologi Informasi, *JUITA : Jurnal Informatika*, **7**(1), 1. https://doi.org/10.30595/juita.v7i1.3491