BAB 2

LANDASAN TEORI

2.1 Emosi

Emosi merupakan suatu kumpulan interaksi yang kompleks antara faktor subjektif dan objektif, dimediasi oleh sistem saraf atau hormonal, yang dapat (a) membangkitkan pengalaman afektif seperti perasaan gairah, kenikmatan/ketidaknikmatan; (b) menghasilkan proses kognitif seperti efek persepsi yang relevan secara emosional, penilaian, proses pelabelan; (c) mengaktivasi penyesuaian psikologis yang luas terhadap kondisi yang menggugah; dan (d) menyebabkan perilaku yang sering, tetapi tidak selalu, ekspresif, diarahkan pada tujuan, dan adaptif (Kleinginna & Kleinginna, 1981).

Selain itu, emosi mempunyai pengertian yang luas dan sempit. Secara sempit, emosi merujuk pada apa yang disebut dengan *full-blown emotion*, dimana emosi secara sementara merupakan fitur dominan dari kehidupan mental, yang mendahului pertimbangan biasa dan sangat mengarahkan orang-orang ke arah tindakan yang dikendalikan oleh emosi. Pengertian luas mencakup apa yang disebut *underlying emotion* atau emosi pokok, yang mewarnai pemikiran dan perilaku seseorang ke tingkat yang lebih besar atau kecil tanpa perlu merebut kendali (Cowie et al., 2001).

Emosi memegang peranan penting dalam komunikasi manusia dan dapat di-

ekspresikan secara verbal melalui kosakata emosional atau melalui isyarat nonverbal seperti intonasi suara, ekspresi wajah, dan gerak-gerik (Koelstra et al., 2012). Secara umum, terdapat enam jenis emosi dasar yaitu marah, takut, sedih, senang, jijik, dan kaget (Ekman, 1992). Berikut adalah beberapa jenis emosi yang akan diklasifikasikan dalam penelitian ini :

1. Emosi Senang

   Emosi senang termasuk di dalam emosi positif dasar (Izard, 2009). Emosi ini dapat dipicu melalui beberapa hal, misalnya ketika seseorang menemukan calon pasangan, dan diartikan sebagai rasa memiliki. Dalam bahasa subjektif, emosi ini disebut senang atau kebahagiaan (Cowie et al., 2001; Plutchik, 1980).

2. Emosi Sedih

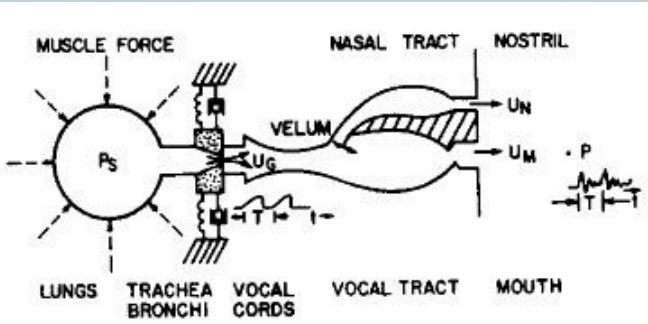
   Emosi sedih merupakan bagian dari emosi negatif dasar (Izard, 2009). Salah satu perangsang terjadinya emosi ini adalah ketika kehilangan seseorang yang berharga dan diartikan sebagai meninggalkan atau ditinggalkan. Perilaku yang umumnya terlihat ketika seseorang merasa sedih adalah menangis (Cowie et al., 2001; Plutchik, 1980).

3. Emosi Netral

   Netral dapat dikategorikan sebagai emosi yang bukan merupakan emosi positif juga bukan emosi negatif, sebagai contohnya adalah tenang (Qiao, Li, Xiang, & Deng, 2011).

2.2 Sinyal Suara Manusia

Suara manusia merupakan gelombang akustik yang diradiasikan dari sistem *sub-glottal* ketika udara dikeluarkan dari paru-paru dan aliran udara yang dihasilkan diusik oleh penyempitan di suatu tempat di saluran vokal. Sistem *sub-glottal* tersebut terdiri atas paru-paru, bronkeas, dan trakea, yang berfungsi sebagai sumber energi untuk produksi suara (Rabiner & Schafer, 1978). Selama manusia berbicara, saluran vokal yang berada di atas pangkal tenggorokan secara terus-menerus mengubah bentuknya, memproduksi pola frekuensi *formant* dengan waktu bervariasi. Frekuensi *formant* merupakan frekuensi akustik yang diijinkan oleh saluran vokal untuk melewatinya dengan peredaman minimum (Lieberman, 2007).



Gambar 2.1 Diagram sistem vokal (Rabiner & Schafer, 1978)

Sinyal suara manusia mengandung tiga daerah yang berbeda menurut cara perangsangan atau eksitasinya (Das, Das, & Bhattacharjee, 2014; Rabiner & Schafer, 1978; Rani, Rani, Ravi, & Sree, 2014) :

1. *Voiced*

Pada daerah bersangkutan, terdapat suara yang dihasilkan dari memaksakan

udara melewati *glottis* dengan tekanan dari pita suara yang disesuaikan agar bergetar dalam gerakan relaksasi.

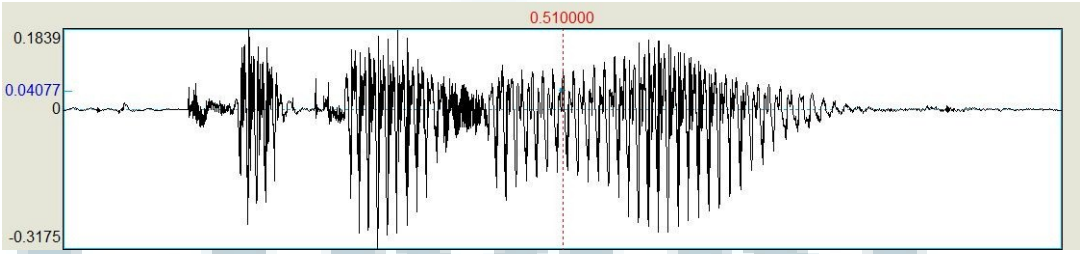
2. *Unvoiced*

   Daerah sinyal yang memiliki suara yang dihasilkan tanpa getaran dari pita suara disebut dengan *unvoiced*. Apabila masukan eksitasi seperti *noise* maka akan menghasilkan sinyal acak seperti *noise* tanpa sifat periodik.

3. *Silenced*

   Pada daerah *silenced*, tidak ada eksitasi yang terjadi pada saluran vokal sehingga tidak ada suara yang dihasilkan.

Suara manusia termasuk sinyal non-stasioner dikarenakan sifat waktu yang bervariasi dan perubahan tiba-tiba dalam frekuensi dari sistem produksi suara manusia. Sinyal non-stasioner dicirikan dengan perubahan yang tidak tetap pada frekuensinya (Yadav et al., 2015).



Gambar 2.2 Sinyal suara manusia mengucapkan *two thousand one*

2.3 Budaya dan Emosi pada Suara

Banyak usaha yang telah dilakukan untuk memahami pengaruh budaya terhadap emosi di dalam suara dan para peneliti telah melaporkan sejumlah penemuan

bahwa terdapat beberapa karakteristik umum yang mengijinkan emosi yang serupa didiskriminasi secara universal antar budaya. Namun terdapat pula atribut suara yang memfasilitasi pengenalan emosi eksklusif dari budaya tertentu (Kamaruddin et al., 2012).

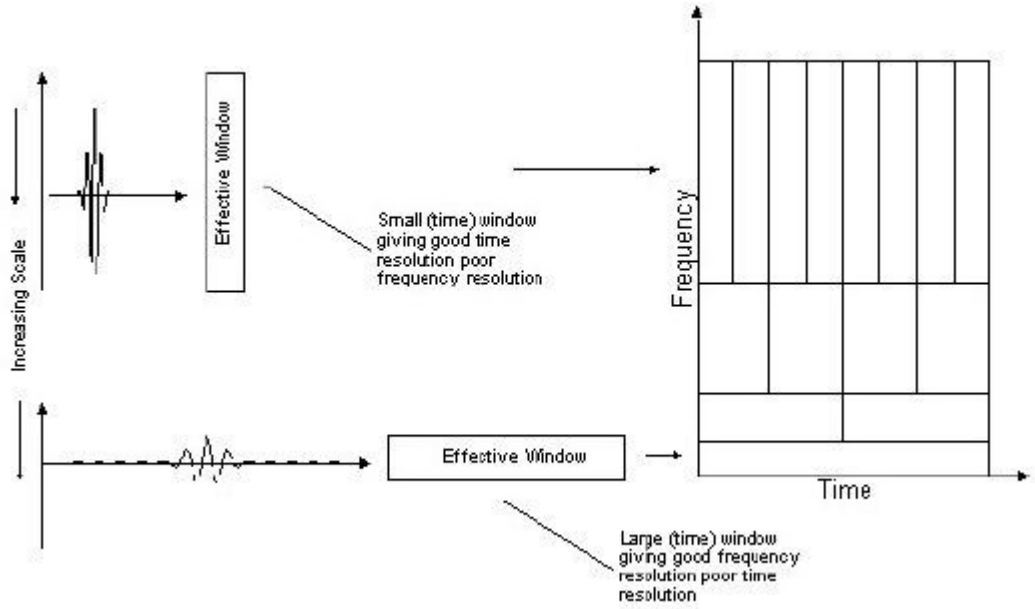
Terdapat penelitian yang bertujuan untuk mengetahui apakah ekspresi vokal dari emosi yang dikeluarkan oleh seorang aktor dari budaya tertentu dapat diidentifikasi oleh anggota dari budaya atau negara berbeda. Kalimat tak berarti yang tersusun atas suku kata dari enam bahasa Eropa diucapkan oleh beberapa aktor berkebangsaan Jerman dalam lima emosi berbeda, lalu digunakan sebagai sampel suara dalam penelitian. Dari sembilan negara yang menjadi partisipan dalam penelitian, Jerman menempati urutan pertama dengan akurasi pengenalan emosi sebesar 74% dan Indonesia pada urutan terakhir dengan akurasi sebesar 52%. Hal tersebut disebabkan oleh meningkatnya ketidakmiripan bahasa meskipun kalimat sampel yang digunakan tidak memiliki arti (Scherer, Banse, & Wallbott, 2001). Dengan ini dapat disimpulkan bahwa bahasa merupakan salah satu parameter kebudayaan dan mempengaruhi performa dari pengenalan emosi (Kamaruddin et al., 2012).

2.4 *Wavelet Transform*

Short-Time Fourier Transform (STFT) sering dimanfaatkan untuk analisis sinyal non-stasioner. STFT menggunakan fungsi jendela dengan lebar konstan, yang digeser menurut waktu. Pada frekuensi tinggi, jumlah gelombang dalam sebuah jendela menjadi tinggi, menghasilkan akurasi yang baik untuk pengukuran frekuensi,

tetapi tidak untuk lokalisasi waktu. Mempersempit lebar jendela dapat mengakomodasi lokalisasi waktu tetapi tidak sesuai untuk frekuensi rendah karena panjang gelombang yang melebihi lebar jendela tidak dapat diukur (Barford, Fazzio, & Smith, 1992).

Solusi dari masalah tersebut adalah untuk mengadopsi strategi jendela yang fleksibel dan secara selektif mengubah sinyal sesuai dengan kebutuhan waktu atau frekuensi. *Wavelet* dapat mengatasi hal tersebut dengan proses penskalaan dan translasi. Penskalaan berarti meregangkan *wavelet* untuk memastikan energi yang dikandung dalam *wavelet* yang diskalakan sama dengan *mother wavelet* yang asli. Translasi menggerakkan *wavelet* ke arah positif, sepanjang sumbu x. Transformasi diperoleh dengan penskalaan dan translasi berkelanjutan dari *mother wavelet* sepanjang sinyal tersebut. Komponen skala tinggi (frekuensi rendah) diregangkan dan memenuhi jendela yang lebih besar, memberikan resolusi waktu yang buruk dan resolusi frekuensi yang baik (serupa pada STFT dengan jendela besar), begitu pula sebaliknya (Leavey et al., 2003).



Gambar 2.3 Skema *windowing* fleksibel dari *wavelet* (Leavey et al., 2003)

2.4.1 *Discrete Wavelet Transform*

Pada *Discrete Wavelet Transform* (DWT), sinyal didekomposisi dengan menggunakan penyaringan *high pass* dan *low pass*. Keluaran dari saringan *low pass* adalah koefisien aproksimasi dan *high pass* adalah koefisien detil (Yadav et al., 2015). Fungsi *wavelet* berperan sebagai saringan *high pass* dan fungsi penskalaan sebagai saringan *low pass*. Proses dekomposisi dilakukan pada koefisien aproksimasi tingkat pertama, lalu pada koefisien aproksimasi tingkat kedua, dan seterusnya sampai tingkat dekomposisi yang diinginkan. Persamaan untuk *Discrete Wavelet Transform* adalah

$$\psi(j,k) = \frac{1}{\sqrt{s_0^j}} \psi\left(\frac{t - k\tau_0 s_0^j}{s_0^j}\right) \quad (1)$$

di mana j dan k adalah bilangan bulat, s_0 adalah langkah dilasi yang tetap, τ_0

adalah faktor translasi. Umumnya, s_0 ditetapkan sebagai 2 dan τ_0 sebagai 1 (Leavey et al., 2003).

2.4.2 Daubechies *wavelet*

Daubechies *wavelet* merepresentasikan dasar dari pemrosesan sinyal *wavelet* dan telah banyak digunakan dalam penelitian untuk mengekstraksi fitur yang dibutuhkan dalam proses klasifikasi (Anusuya & Katti, 2011; Sunny, Peter, & Jacob, 2011; Tzanetakis, Essl, & Cook, 2001). Keluarga *wavelet* Daubechies bersifat *orthogonal* dan mempunyai *vanishing moment* yang terbatas. Untuk *wavelet* Daubechies dengan orde yang lebih tinggi ψ_{dbN}, N menandakan orde dari wavelet dan jumlah dari *vanishing moment* (Daubechies, 1992; Salem, Ghamry, & Meffert, 2009).

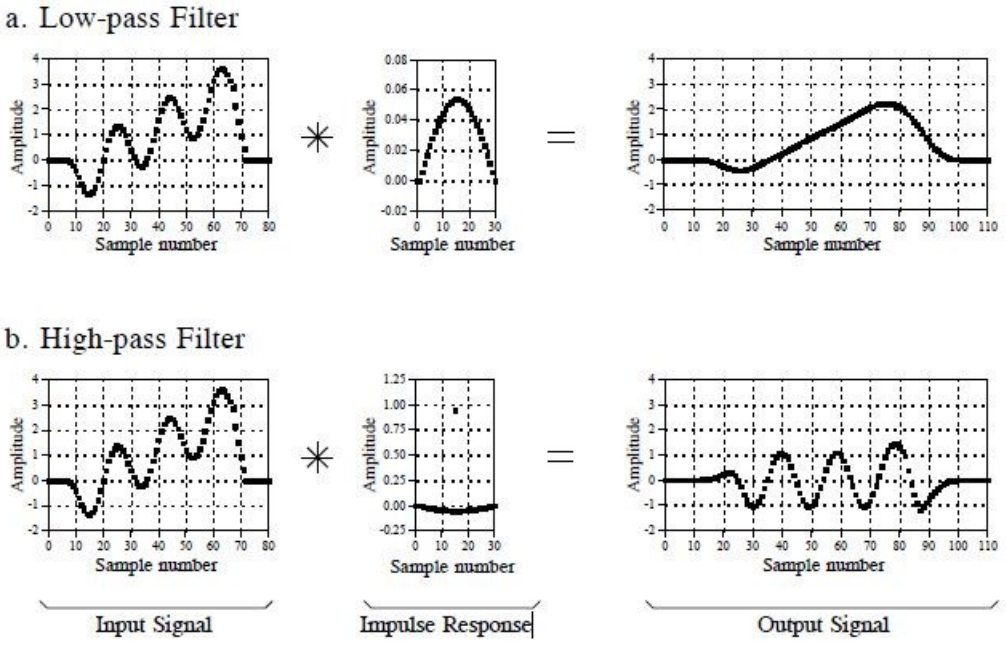
2.5 Konvolusi

Konvolusi adalah cara matematis dari mengkombinasikan dua sinyal untuk menghasilkan sinyal ketiga. Sebuah sinyal masukan, $x[n]$, memasuki sistem linear dengan sebuah respon impuls, $h[n]$, menghasilkan sinyal keluaran, $y[n]$. Dalam bentuk persamaan, konvolusi direpresentasikan oleh simbol bintang (*).

$$x[n] * h[n] = y[n] \tag{2}$$

Salah satu pemanfaatan konvolusi adalah untuk penyaringan *high-pass* dan

low-pass. Pada (a) , respon impuls untuk saringan *low-pass* berbentuk lengkungan halus, menghasilkan hanya gelombang landai yang berubah dengan lambat yang dilewatkan sebagai keluaran. Untuk saringan *high-pass* (b), mengijinkan hanya gelombang sinus yang berubah dengan cepat untuk lewat (Smith, 1997).



Gambar 2.4 Konvolusi pada penyaringan *high-pass* dan *low-pass* (Smith, 1997)

2.6 *Linear Mixed Effect Models*

Mixed effect models, seperti tipe model statistik yang lainnya, menjelaskan hubungan antara variabel respon dengan beberapa kovariat yang diukur atau diobservasi beserta respon tersebut (Bates, 2010). Model ini merujuk pada model yang memiliki fitur kunci yaitu *fixed* dan *random effect*. *Linear mixed effect models* merupakan model yang secara sederhana memodelkan *fixed* dan *random effect* sebagai

bentuk linear. Bentuk dari *linear mixed effect model* adalah

$$y_{ij} = \beta_1 x_{1ij} + \beta_2 x_{2ij} ... \beta_n x_{nij} + b_{i1} z_{1ij} + b_{i2} z_{2ij} ... b_{in} z_{nij} + \varepsilon_{ij} \quad (3)$$

di mana y_{ij} merupakan nilai dari variabel hasil untuk kasus ij tertentu, β_1 sampai β_n merupakan koefisien *fixed effect* (seperti koefisien regresi), x_{1ij} sampai x_{nij} adalah variabel *fixed effect* untuk observasi j dalam kelompok i, b_{i1} sampai b_{in} adalah koefisien *random effect* yang diasumsikan sebagai multivariat terdistribusi normal, z_{1ij} sampai z_{nij} merupakan variabel *random effect*, dan ε_{ij} adalah *error* untuk kasus j dalam kelompok i di mana *error* setiap kelompok diasumsikan sebagai multivariat terdistribusi normal (Starkweather, 2010).

Fixed effect pada *mixed effect models* merupakan kovariat yang mempunyai tingkatan yang tetap dan dapat digandakan. Jika tingkatan dari kovariat yang diobservasi merupakan sampel acak dari seluruh tingkatan yang mungkin, maka kovariat dimasukkan sebagai *random effect* di dalam model (Bates, 2010). *Fixed effect* menggambarkan perubahan respon di antara tiap tingkatan kovariat, misalnya perubahan denyut jantung pasien pada menit pertama, menit kelima, dan lain-lain. *Random effect* menggambarkan variasi yang diinduksi pada respon oleh tingkatan berbeda dari kovariat, misalnya variasi denyut jantung antar pasien (Bates, 2005).

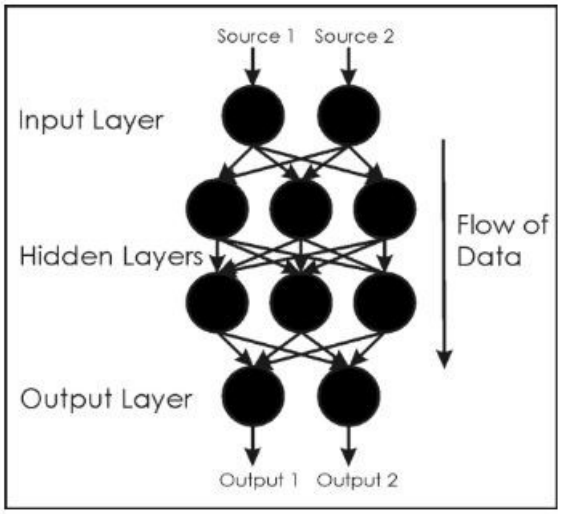
2.7 *Likelihood Ratio Test*

Likelihood ratio test merupakan uji yang membandingkan nilai dari fungsi kemungkinan untuk dua model bercabang yang mendefinisikan hipotesis yang diuji. Hipotesis tersebut dapat mengenai parameter *fixed effect* atau kovarians dalam *linear mixed effect model*. Pada umumnya, pengujian ini mengharuskan bahwa kedua model dipersiapkan menggunakan bagian data yang sama (Østgård, 2011). Contoh penerapan dari *likelihood ratio test* adalah pengujian untuk mengetahui apakah *random effect* tertentu diperlukan di dalam model. Dalam bahasa pemrograman R, uji tersebut dapat dilakukan menggunakan fungsi *anova* (Baayen, Davidson, & Bates, 2008). Jika statistik hasil uji signifikan pada tingkat α, yaitu $p < \alpha$, maka hipotesis nol yang menyatakan *random effect* tidak diperlukan dapat ditolak dan *random effect* dipertahankan di dalam model (Østgård, 2011).

2.8 *Neural Network*

Neural network atau *artificial neural network* (ANN) adalah sistem biologis yang mendeteksi pola dan membuat prediksi (Andrews, Diederich, & Tickle, 1995; Jain & Srivastava, 2013). *Neural network* terdiri atas banyak neuron di mana neuron merupakan unit dasarnya (Medhat, Hassan, & Korashy, 2014). Salah satu kegunaan dari *neural network* adalah untuk mengklasifikasikan atau mengkategorikan sesuatu. Jika gambar dari sekelompok orang disimpan di dalam jaringan, *neural network* dapat menyimpulkan siapa yang terlihat sedih dan siapa yang terlihat senang (Wi-

lde, 2013).



Gambar 2.5 *Neural network* dengan lapisan tersembunyi (Jain & Srivastava, 2013)

Neural network untuk masalah klasifikasi dapat dilihat sebagai fungsi pemetaan, $F : R_d \rightarrow R_M$, di mana masukan x berdimensi d diserahkan ke dalam jaringan dan output jaringan y bervektor M diperoleh untuk membuat keputusan klasifikasi. Jaringan dibangun sedemikian rupa agar pengukuran kesalahan keseluruhan seperti *mean squared error* (MSE) diminimalisir (Zhang, 2000). Terdapat dua jenis algoritma pembelajaran yang dapat digunakan untuk melatih *neural network* (Ruiz & Srinivasan, 1998):

1. *Supervised Learning*

   Masukan untuk jaringan adalah sekelompok sampel yang meliputi fitur masukan dan keluaran yang diekspektasikan untuk tiap contoh. Algoritma bersangkutan disebut *supervised* karena selama fase pelatihan, bobot dari jaringan disesuaikan sampai keluarannya mendekati keluaran yang diinginkan.

2. *Unsupervised Learning*

   Dalam algoritma *unsupervised*, masukan hanya berupa nilai dari fitur dan jaringan akan melakukan pengelompokan atau prosedur asosiasi untuk mempelajari kelas yang ada di dalam data pelatihan.

Neural network memiliki beberapa kelebihan sebagai alat untuk klasifikasi, antara lain (Zhang, 2000):

1. Dapat menyesuaikan diri dengan data tanpa perlu bentuk spesifikasi fungsional atau distribusional eksplisit dari model pokok,
2. Dapat mengaproksimasi fungsi dengan akurasi berubah-ubah,
3. Merupakan model non-linear yang dapat dengan fleksibel memodelkan hubungan kompleks di dunia nyata,
4. Dapat mengestimasi probabilitas posterior yang menyediakan dasar untuk menciptakan peraturan klasifikasi dan melakukan analisis statistik.

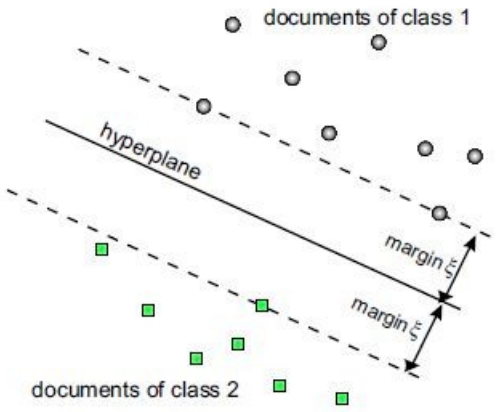
2.9 *Support Vector Machine*

Sebuah *Support Vector Machine* (SVM) hanya dapat memisahkan dua kelas yaitu kelas positif L_1 (diindikasikan dengan $y = +1$) dan kelas negatif L_2 (diindikasikan dengan $y = -1$). Dalam ruang vektor masukan, sebuah *hyperplane* dapat

didefinisikan dengan mengatur $y = 0$ dalam persamaan linear berikut.

$$y = f(\overrightarrow{t_d}) = b_0 + \sum_{j=1}^{n} b_j t_{dj} \tag{4}$$

Algoritma SVM menentukan *hyperplane* yang diletakkan di antara kelas positif dan negatif. Parameter b_j diadaptasikan sedemikian rupa sehingga jarak ξ di antara *hyperplane* dengan contoh positif dan negatif dimaksimalkan, seperti pada Gambar 2.6. Contoh yang memiliki jarak ξ dari *hyperplane* disebut *support vector* dan menentukan lokasi *hyperplane* yang sebenarnya. Contoh dengan vektor $\overrightarrow{t_d}$ dikategorikan ke dalam L_1 jika nilai dari $f(\overrightarrow{t_d}) > 0$ dan L_2 jika sebaliknya. Dalam kasus vektor contoh dari dua kelas tidak dapat dipisahkan secara linear, *hyperplane* dipilih sehingga sesedikit mungkin contoh diletakkan di sisi yang salah (Hotho et al., 2005). Karena SVM menggunakan perlindungan *overfitting*, yang tidak selalu bergantung pada jumlah fitur, SVM memiliki potensi untuk menangani ruang fitur yang besar dan mengeliminasi kebutuhan pemilihan fitur (Joachims, 1998).



Gambar 2.6 *Hyperplane* dengan jarak maksimal ke contoh positif & negatif yang dibentuk oleh SVM (Joachims, 1998)

2.10 *Principal Component Analysis*

Principal Component Analysis (PCA) merupakan teknik yang menggunakan prinsip matematis yang mutakhir untuk mengubah sejumlah variabel yang mungkin berkorelasi ke dalam jumlah lebih kecil yang disebut dengan *principal components* (Richardson, 2009). *Principal components* adalah kumpulan variabel yang tidak berkorelasi, di mana variabel ini merupakan kombinasi linear dari variabel asli dan diurutkan menurun menurut kepentingan sehingga *principal component* pertama menerangkan sebanyak mungkin variasi dalam data asli (Chatfield & Collins, 1980). Secara umum, PCA menggunakan transformasi ruang vektor untuk mengurangi dimensionalitas dari kumpulan data. Dengan data yang telah dikurangi dimensinya, mengijinkan pengguna untuk menemukan tren, pola dan *outlier* lebih mudah daripada tidak melakukan PCA (Richardson, 2009).

2.11 *K-Fold Cross Validation*

Dalam *k-fold cross validation*, kumpulan data secara acak dibagi ke dalam k bagian (*fold*) D_1, D_2, D_3 dengan ukuran yang kurang lebih sama. Algoritma yang membangun *classifier* dilatih dan diuji sebanyak k kali, setiap waktu $t \in \{1, 2, ..., k\}$, dilatih dengan $D \backslash D_t$ dan diuji dengan D_t. Estimasi akurasi *cross validation* adalah jumlah keseluruhan dari klasifikasi yang benar dibagi dengan jumlah data. Secara formal, D_i adalah kumpulan data uji yang memuat data $x_i = \langle v_i, y_i \rangle$, maka estimasi

akurasi *cross validation* dapat diperoleh dengan persamaan

$$acc_{cv} = \frac{1}{n} \sum_{\langle v_i, y_i \rangle \in D} \delta(I(D \backslash D_{(i)}, v_i), y_i) \tag{5}$$

K-fold cross-validation dengan nilai *k* sedang (10-20) mengurangi variasi akurasi dan meningkatkan *bias*. Salah satu jenis dari *k-fold cross validation* adalah *stratified k-fold cross validation*. Dalam jenis ini, tiap *fold* memuat proporsi data yang kurang lebih sama untuk tiap kelas. Menurut penelitian terdahulu, *stratified* 10-*fold cross validation* direkomendasikan untuk pemilihan model (Kohavi, 1995).

2.12 Friedman *Test*

Uji Friedman digunakan untuk data ordinal dalam situasi uji hipotesis yang melibatkan dua atau lebih sampel yang berhubungan. Uji ini merupakan alternatif ketika satu atau lebih asumsi dari ANOVA pengukuran berulang tidak terpenuhi. Jika hasil dari uji Friedman adalah signifikan, hal tersebut mengindikasikan bahwa terdapat perbedaan signifikan di antara paling sedikit dua median sampel dari kumpulan *k* median (Sheskin, 2003). Akan tetapi, uji Friedman tidak mengidentifikasikan di mana dan berapa perbedaan yang terjadi. Untuk mengetahui perbedaan tertentu antara pasangan sampel dapat digunakan uji *post hoc*, seperti uji Wilcoxon.

Untuk menghitung statistik uji Friedman F_r, data disusun dalam bentuk tabel, di mana subjek sebagai baris dan nilai dari tiap kondisi sebagai kolom. Kemudian, berikan ranking pada nilai subjek untuk setiap kondisi. Jika tidak terdapat ranking

yang sama, rumus berikut digunakan untuk menentukan F_r:

$$F_r = \left[\frac{12}{nk(k+1)} \sum_{i=1}^{k} R_i^2 \right] - 3n(k+1) \tag{6}$$

di mana n adalah jumlah baris atau subjek, k adalah jumlah kolom atau kondisi, dan R_i adalah jumlah ranking dari kolom atau kondisi i. Jika terdapat ranking yang sama, rumus berikut yang digunakan :

$$F_r = \frac{n(k-1)\left[\sum_{i=1}^{k} \frac{R_i^2}{n} - C_F\right]}{\sum r_{ij}^2 - C_F} \tag{7}$$

di mana C_F adalah perbaikan untuk nilai yang sama, $\frac{1}{4}nk(k+1)^2$, dan r_{ij} adalah ranking subjek j di kolom i. Statistik F_r dapat dibandingkan dengan tabel yang mempunyai distribusi X^2 untuk mencari nilai kritis yang digunakan untuk memeriksa perbedaan signifikan pada kelompok (Corder & Foreman, 2014).

2.13 Wilcoxon *Signed-Rank Test*

Pengujian dua sampel berhubungan digunakan untuk menguji apakah dua sampel yang berpasangan berasal dari populasi yang sama. Jika benar demikian, maka ciri-ciri kedua sampel seperti rata-rata, median, dan lainnya relatif sama untuk kedua sampel maupun populasinya. Uji Wilcoxon merupakan uji dua sampel berhubungan yang digunakan sebagai alternatif dari uji *t-paired* ketika salah satu dari syarat uji parametrik tidak terpenuhi yaitu (1) data bertipe ordinal atau nomi-

nal, (2) data bertipe interval atau rasio, namun tidak berdistribusi normal (Santoso, 2010).

Rumus untuk memperoleh nilai z dari uji Wilcoxon adalah :

$$Z^* = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{8}$$

di mana T adalah nilai terkecil di antara jumlah ranking dengan perbedaan positif ($\sum R_+$) dan jumlah ranking dengan perbedaan negatif ($\sum R_-$), n adalah jumlah pasangan dalam analisis. Nilai ini akan dibandingkan dengan tabel *z-score* untuk memperoleh nilai kritis yang dapat digunakan untuk memeriksa perbedaan signifikan antar kelompok (Corder & Foreman, 2014).

2.14 Praat

Praat adalah perangkat lunak untuk analisis suara yang tersedia secara gratis dengan kode *open source* yang dikembangkan oleh Paul Boersma dan David Weenink (Styler, 2013). Praat memiliki banyak fungsionalitas di antaranya adalah perekaman suara, analisis sinyal suara, pelabelan dan segmentasi, sintesis suara, eksperimen mendengar, manipulasi suara, algoritma pembelajaran (contohnya: *neural network*), statistik, dan lain-lain (Boersma & Weenink, 2009).

2.15 Penelitian Terdahulu

2.15.1 *Speech Emotion Recognition Using Support Vector Machine*

Pengenalan emosi melalui suara telah banyak menggunakan fitur prosodi dan spektral. Dalam penelitian ini, dibandingkan persentase pengenalan emosi berdasarkan beberapa fitur dan kombinasinya. Suara emosional untuk penelitian diperoleh dari *Berlin German Database* dan SJTU *Chinese Database* yang dibangun sendiri oleh peneliti karena kurangnya *database* dalam bahasa Mandarin.

Seperti sistem pengenalan pola pada umumnya, sistem pengenalan emosi melalui suara ini mengandung empat modul utama, yaitu masukan suara, ekstraksi fitur, klasifikasi dengan SVM, dan keluaran emosi. Adapun fitur yang diekstraksi adalah :

1. Statistik dari energi di dalam keseluruhan sampel suara seperti nilai rata-rata, maksimum, variansi, jarak variasi, kontur dari energi.
2. Statistik dari *pitch* seperti nilai rata-rata, variansi, jarak variasi dan kontur.
3. *Linear Prediction Cepstrum Coefficients* (LPCC) yang diperoleh dari komputasi perulangan *Linear Prediction Coefficients* (LPC) sesuai dengan *all-pole model*.
4. *Mel-Frequency Cepstrum Coefficients* (MFCC) di mana yang diekstraksi adalah 12 koefisien pertama.

5. *Mel Energy Spectrum Dynamic Coefficients* (MEDC) dengan proses yang serupa dengan MFCC. Perbedaannya adalah MEDC mengambil rata-rata logaritma dari energi setelah *Mel filter bank* dan pengemasan frekuensi. Perbedaan pertama dan kedua juga dihitung dari fitur ini.

Model pelatihan untuk *support vector machine* dibangun dari kombinasi fitur yang berbeda dan dianalisis akurasi pengenalannya. Jumlah sampel yang diperoleh dari *database* Berlin untuk emosi sedih, senang, dan netral adalah masing-masing 62, 71, dan 79 sampel suara, dan dari *database* berbahasa Mandarin sebanyak 500 sampel per emosi.

Training Model	*Combination of Feature Parameters*
Model1	*Energy+Pitch*
Model2	*MFCC+MEDC*
Model3	*MFCC+MEDC+LPCC*
Model4	*MFCC+MEDC+Energy*
Model5	*MFCC+MEDC+Energy+Pitch*

Gambar 2.7 Kombinasi berbeda dari fitur suara (Pan et al., 2012)

Libsvm dari Matlab dimanfaatkan untuk melakukan validasi silang dari model pelatihan dan analisis hasil. Untuk setiap emosi, sampel suara dibagi menjadi 90% untuk pelatihan dan 10% untuk pengujian. Dari hasil yang diperoleh, kombinasi fitur terbaik untuk *database* Berlin adalah MFCC + MEDC + Energi, karena persentase validasinya dapat mencapai 95% dan akurasi 91,3%. Ketika ditambahkan fitur LPCC, performa model menurun yang mungkin dihasilkan dari redundansi fitur. Untuk *database* berbahasa mandarin, MFCC + MEDC + Energi juga menunjukkan performa yang baik dengan persentase validasi silang 95% dan akurasi 95%.

Kombinasi ini mempunyai performa lebih baik pada *database* berbahasa Mandarin yang berarti energi mempunyai peranan penting dalam pengenalan emosi pada suara berbahasa Mandarin.

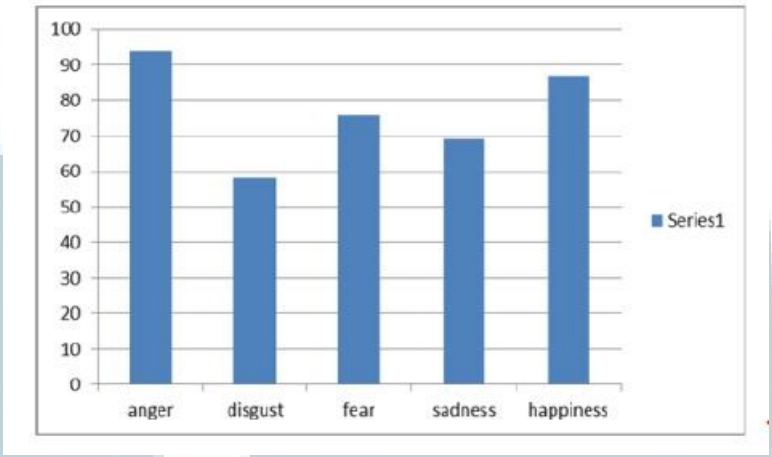
Dapat disimpulkan bahwa kombinasi berbeda dari fitur karakteristik emosional dapat menghasilkan persentase pengenalan emosi yang berbeda, dan sensitivitas dari fitur emosi dalam bahasa lain juga berbeda sehingga perlu penyesuaian fitur untuk bermacam-macam *database*.

2.15.2 *Emotion Recognition from Persian Speech with Neural Network*

Penelitian ini bertujuan untuk melakukan pengenalan emosi dari suara berbahasa Persia, dikarenakan adanya perbedaan struktur bahasa dan budaya yang menyebabkan perbedaan pengenalan emosi dari satu bahasa ke bahasa lainnya. Untuk penelitian ini, *database* suara emosional berbahasa Persia yang tepat belum tersedia, maka *database* tersebut dibangun dari ucapan aktor dan aktris yang diperoleh dari 60 film berbeda, karena film mengekspresikan emosi yang lebih alami. *Database* ini berisi lebih dari 2400 suara emosional berbahasa Persia dalam enam mode yaitu marah, jijik, takut, sedih, senang, dan normal. Ucapan direkam dengan menggunakan perangkat lunak Praat dan panjangnya berbeda satu sama lain.

Fitur yang diekstraksi dari sinyal suara adalah energi, *pitch*, koefisien MFCC, turunan pertamanya, dan juga titik maksimum dan minimum, serta data normalisasi yang dihitung dari suara normal pembicara. Klasifikasi emosi kemudian dilakukan dengan *multi-layer perceptron neural network*, tersusun atas lapisan masukan,

lapisan tengah, dan lapisan keluaran. Masukan jaringan berupa vektor fitur dan keluarannya berupa label kelas emosi. Jaringan ini dilatih dengan menggunakan masukan dan label biner.



Gambar 2.8 Hasil klasifikasi tiap emosi (Hamidi & Mansoorizade, 2012)

Penelitian ini berhasil mengklasifikasikan emosi dengan akurasi rata-rata 78%. Namun, masih terdapat beberapa persoalan untuk penelitian dalam bidang ini seperti tersedianya *database* untuk masing-masing bahasa, lingkungan dalam *database* (bising atau tanpa bising), ucapan yang alami atau disimulasikan, jumlah emosi di dalam *database*, serta tipe fitur yang digunakan.

2.15.3 *Comparison of Different Speech Feature Extraction Techniques with and without Wavelet Transform to Kannada Speech Recognition*

Pada penelitian berikut, dibandingkan proses ekstraksi fitur yang dapat meningkatkan akurasi pada sistem pengenalan kata berbahasa Kannada. Teknik *wavelet transform* pertama adalah DWT di mana dekomposisi tingkat berikutnya dilaku-

kan pada sinyal *low pass* yang diperoleh dari dekomposisi sebelumnya. Keluarga *wavelet* yang digunakan untuk tujuan eksperimen adalah DB8-tingkat 5 karena merupakan yang terbaik di antara keluarga Daubechies. *Wavelet packet decomposition* (WPD) merupakan teknik kedua di mana dekomposisi dilakukan juga pada koefisien detil, menawarkan analisis yang lebih mendalam. Dekomposisi WPD juga dilakukan sampai tingkat 5 dengan Shannon *entropy*. Selain untuk ekstraksi fitur, teknik ini digunakan untuk membuang bising dari suara.

Sebanyak 10 wanita dewasa yang berbahasa Kannada diminta mengucapkan 10 kata Kannada yaitu angka satu sampai sepuluh sebanyak 10 kali per kata, disimpan dalam format Microsoft *wave* dengan *sampling rate* 8000 Hz, 16 bit PCM dan kanal mono. Perangkat lunak yang digunakan adalah Praat dan sinyal suara yang dikumpulkan terdiri atas sinyal bersih dan bising.

Number	Kannada Word	Symbol used in the paper
1	"Ondu"	One
2	"Eradu"	Two
3	"Muru"	Three
4	"Nalaku"	Four
5	"Iydu"	Five
6	"Aaru"	Six
7	"Yelu"	Seven
8	"Enttu"	Eight
9	"Ombatthu"	Nine
10	"Hatthu"	Ten

Gambar 2.9 Daftar kata dalam Bahasa Kannada (Anusuya & Katti, 2011)

Dari 1000 sampel suara yang dikumpulkan, didekomposisi ke dalam koefisien DWT atau WPD menurut jenis sinyalnya. Keduanya dihitung sampai de-

komposisi tingkat lima. Kemudian, hasil dekomposisi disimpan dari urutan terkecil sampai terbesar, mulai dari 10 sampel untuk angka satu, dan seterusnya. Keseluruhan sampel ini digunakan untuk tahap pelatihan. Untuk tahap pengujian, 500 sampel dikumpulkan terpisah yang terdiri atas sinyal bersih dan bising dari 10 pembicara wanita mengucapkan 10 kata sebanyak 5 kali per kata.

Berikut beberapa koefisien yang dibutuhkan dihitung dari koefisien DWT dan WPD untuk masing-masing sinyal bersih dan bising :

1. Sepuluh koefisien LPC pertama.
2. Tiga belas koefisien MFCC pertama.
3. Lima koefisien *autocorrelation* pertama untuk memperoleh koefisien RASTA-PLP.

Nilai rata-rata koefisien dari masing-masing kata dikalkulasi lalu dibandingkan dengan sinyal uji menggunakan Euclidian *distance*. Dari 10 kelompok (kata satu sampai sepuluh), sinyal uji dengan jarak paling kecil dideklarasikan sebagai sinyal teridentifikasi. Sinyal uji sebanyak 5 sampel berbeda per kata diperoleh dari *database* uji yang dibuat sebelumnya.

Hasil penelitian menunjukkan bahwa fitur yang diekstraksi dari *wavelet transform* mempunyai akurasi paling tinggi untuk pengenalan kata. Kombinasi fitur yang diperoleh dari RASTA-PLP dan *wavelet* memberikan hasil yang lebih baik pada sinyal bising, sedangkan kombinasi fitur LPC atau MFCC dengan koefisien *wavelet* memberikan hasil yang baik untuk sinyal bersih. Secara khusus, dekomposisi de-

ngan Daubechies 8 tingkat 5 menghasilkan kesuksesan dengan persentase tertinggi untuk pengenalan suara berbahasa Kannada.

2.15.4 *Parametric Speech Emotion Recognition Using Neural Network*

Penelitian ini bertujuan untuk mengklasifikasikan suara yang direkam ke dalam empat kategori : senang, sedih, marah, dan netral menggunakan MATLAB. Sebelum fitur suara diekstraksi, beberapa langkah dilakukan untuk memanipulasi sinyal suara seperti *sampling*, normalisasi, dan segmentasi.

Sampling dilakukan dengan mengumpulkan titik dari sinyal analog pada tingkat T_s tertentu, dengan frekuensi umum untuk sinyal suara adalah 8000Hz, 16000Hz, dan 44100Hz. Di MATLAB, *sampling* secara otomatis diaplikasikan setelah fungsi perekaman. Normalisasi dilakukan untuk memastikan setiap rekaman mempunyai tingkat volume yang sama dengan membagi urutan sinyal dengan nilai maksimum sinyal tersebut. Persamaan normalisasi dapat dilihat di bawah ini.

$$\widetilde{s}(n) = \frac{s(n)}{s_{max}}, n = 1, 2, .., N \tag{9}$$

$s(n)$ merupakan sinyal asli, (n) merupakan sinyal setelah normalisasi, s_{max} adalah nilai maksimum absolut dari urutan sinyal, dan N panjang dari urutan tersebut. Sinyal suara mempunyai karakteristik yang berubah seiring waktu, maka sinyal disegmentasikan ke dalam banyak *frame* yang saling tumpang tindih dengan durasi 10-30ms di mana sinyal diasumsikan stabil.

Fitur yang diekstraksi dari sinyal suara adalah sebagai berikut:

1. Kecepatan berbicara mempunyai hubungan yang kuat dengan emosi seperti senang, sedih, dan marah, ditunjukkan dengan persamaan

$$S_r = \frac{t_v}{n_w} \tag{10}$$

di mana S_r adalah kecepatan berbicara, t_v total durasi dari suara, dan n_w jumlah kata di dalam kalimat.

2. Energi memegang peranan penting dalam pengenalan emosi, dihitung melalui persamaan ini setelah normalisasi.

$$E_n = \sum N^{i=1} x(n)^2 \tag{11}$$

E_n merupakan energi, $x(n)$ urutan sinyal setelah *sampling*, N adalah urutan dari panjang.

3. *Pitch* dikenal sebagai naik dan turunnya nada suara, diperoleh melalui *autocorrelation*.

$$R(k) = \sum FL^{i=1} x(m)x(m+k) \tag{12}$$

Di mana FL adalah panjang frame, $x(m)$ adalah *frame* dari sinyal, k adalah parameter pergeseran, dan $R(k)$ merepresentasikan hasil dari *autocorrelation*.

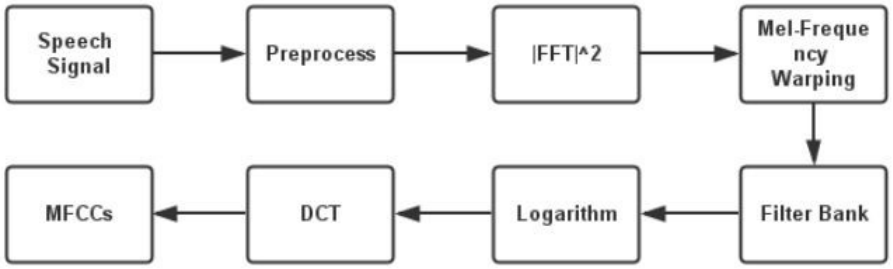
4. Frekuensi *formant* merupakan resonansi di dalam saluran vokal dan menentukan karakteristik warna nada dari huruf hidup. Dapat diperoleh dengan

menghitung akar dari polinomial *Linear Prediction Coding* (LPC).

$$\widetilde{x}[n] = \sum k = 1^{p} a_k x[n-1] \tag{13}$$

$\widetilde{x}[n]$ adalah sampel prediksi, p jumlah dari sampel lampau, dan a_k adalah faktor koefisien.

5. *Mel-Frequency Cepstrum Coefficients* (MFCC) merupakan representasi akurat dari spektrum daya jangka pendek dari sebuah suara. Hanya 13 koefisien pertama yang digunakan.



Gambar 2.10 Proses ekstraksi MFCC (Ma, 2014)

Sampel suara yang digunakan berupa 15 kalimat berbeda untuk empat emosi yang direkam menggunakan MATLAB dengan frekuensi 8000Hz dan durasi 7 detik. *Neural network* untuk klasifikasi memiliki jumlah neuron di lapisan tersembunyi sebanyak 10. Data masukan berupa matriks 15 fitur dari 60 kalimat. Data target adalah matriks 4 emosi dari 60 kalimat, contohnya 1000 untuk marah, 0100 untuk senang, 0010 untuk sedih, dan 0001 untuk netral. Jumlah data untuk pelatihan adalah sebanyak 70 persen, dan 15 persen masing-masing untuk validasi serta pengujian.

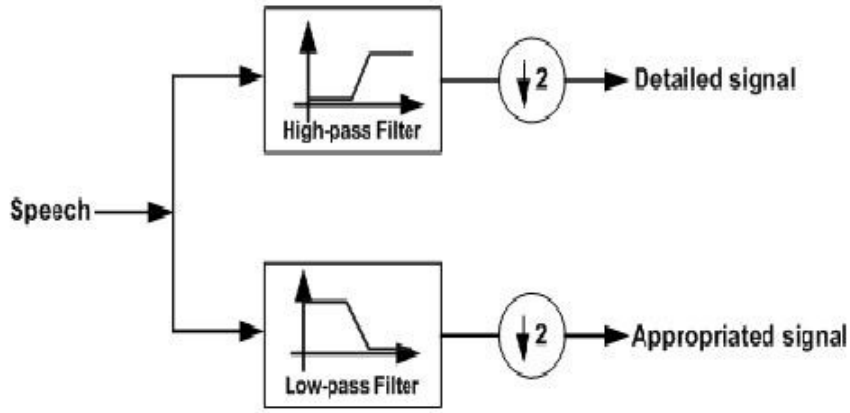
Setelah melalui proses pelatihan sebanyak enam kali, data ucapan baru digunakan sebagai masukan yang akan diklasifikasi, di mana untuk setiap emosi terdapat 10 kalimat. Hasil menunjukkan bahwa 9 ucapan marah, 8 ucapan senang, 7 ucapan sedih, dan 9 ucapan netral diklasifikasikan secara benar. Total persentase klasifikasi untuk data baru ini adalah sebesar 82,5%. Kelemahan dari penelitian ini adalah data suara yang direkam hanya berasal dari satu orang yaitu peneliti sendiri sehingga suara yang dihasilkan untuk masing-masing emosi cenderung sama, misalnya semua rekaman suara sedih terdengar lambat, depresi, dan tidak bertenaga.

2.15.5 *Recognition of Emotion from Marathi Speech Using MFCC and DWT Algorithms*

Joshi and Zalte (2013) mengajukan sistem pengenalan emosi dari sampel suara berbahasa Marathi. *Database* suara terdiri atas kalimat berkelanjutan yang diucapkan dalam enam emosi berbeda (senang, sedih, marah, netral, takut, dan bosan) oleh 10 pembicara (5 wanita dan 5 pria). Sebanyak 10 kalimat diucapkan sebanyak tiga kali per pembicara untuk setiap emosi. Perekaman dilakukan menggunakan mikrofon dengan perangkat lunak Praat, *sampling rate* 16kHz/16 bit, dan jarak antara mulut dan mikrofon disesuaikan sejauh 30 cm.

Sampel suara kemudian dievaluasi oleh 10 pendengar dengan kelompok umur 18 sampai 28 tahun, di mana kalimat dengan emosi yang diidentifikasi oleh paling sedikit 80% dari pendengar yang dipilih untuk penelitian ini. Dari sampel tersebut, 30 sampel per emosi digunakan untuk pelatihan dan 10 sampel untuk tujuan pengu-

jian. Fitur suara diekstraksi dengan algoritma MFCC dan *Discrete Wavelet Transform* (DWT) dengan dekomposisi tingkat empat menggunakan *wavelet* db4. Statistik umum seperti deviasi standar, rata-rata, dan lain-lain digunakan untuk membentuk vektor fitur dari koefisien *wavelet*. Setiap fitur disimpan di *database* dengan label kelas emosinya.



Gambar 2.11 *Discrete wavelet transform* (Joshi & Zalte, 2013)

Algoritma klasifikasi yang dimanfaatkan dalam penelitian ini adalah *Support Vector Machine* (SVM). SVM dibangun dengan memetakan pola pengujian ke ruang fitur dengan dimensi lebih tinggi, di mana titik-titik dapat dipisahkan dengan menggunakan *hyperplane*. Meskipun merupakan *classifier* biner, SVM juga dapat digunakan untuk mengklasifikasikan banyak kelas. Dalam pendekatan SVM, tujuannya adalah memperoleh fungsi $f(x)$ yang menentukan batasan keputusan atau *hyperplane*. Fungsi optimal dari *hyperplane* adalah sebagai berikut.

$$f(x) = \sum_{k=1}^{t} y_k a_k \langle x_k, x \rangle + b \tag{14}$$

Jika masukan titik data x_k mempunyai pengali Langrange (a_k) yang tidak bernilai nol, x_k ini disebut sebagai *support vector*.

Emotion	Emotion Recognition %					
	Anger	Boredom	Fear	Happy	Sad	Neutral
Angry	75	0	0	0	0	25
Boredom	0	84	0	0	16	0
Happy	16	0	16	67	0	0
Sad	0	0	0	11	89	0
Neutral	0	12	12	12	0	64

Gambar 2.12 *Persentase pengenalan dari SVM classifier* (Joshi & Zalte, 2013)

Gambar di atas menunjukkan persentase pengenalan dari masing-masing emosi menggunakan model SVM. Hasil eksperimen ini menunjukkan performa sistem untuk database suara emosional berbahasa Marathi cukup menjanjikan.

Tabel 2.1 Tabel Perbandingan Penelitian Terdahulu

No.	Peneliti (Tahun)	Judul Penelitian	Objek	Metode	Hasil
1	Y. Pan, P. Shen, L. Shen (2012)	*Speech Emotion Recognition Using Support Vector Machine*	Suara manusia berbahasa Jerman dan Mandarin dalam emosi senang,sedih, dan netral.	(1) Kombinasi fitur suara berupa energi, *pitch*, LPCC, MFCC, MEDC. (2) Klasifikasi dengan SVM.	(1) Kombinasi fitur terbaik adalah MFCC + MEDC + Energi dengan akurasi 91,3% pada *database* Jerman dan 95% pada database *Mandarin*. (2) Energi mempunyai peranan penting dalam pengenalan emosi pada suara berbahasa Mandarin.
2	M. Hamidi & M. Mansoorizade (2012)	*Emotion Recognition from Persian Speech with Neural Network*	Suara manusia berbahasa Persia dalam emosi marah, jijik, takut, sedih, senang, dan normal.	(1) *Database* dibangun dari ucapan aktor dan aktris dari 60 film. (2) Ekstraksi fitur berupa energi, *pitch*, koefisien MFCC, titik maks dan min, serta data normalisasi suara normal pembicara. (3) Klasifikasi dengan *neural network*.	Akurasi rata-rata klasifikasi emosi adalah 78%.

3	M.A. Anusuya & S.K. Katti (2011)	*Comparison of Different Speech Feature Extraction Techniques with and without Wavelet Transform to Kannada Speech Recognition*	Suara manusia mengucapkan kata berbahasa Kannada.	(1) Ekstraksi fitur dari koefisien *wavelet* yang diperoleh dari DWT db8-tingkat 5 dan WPD tingkat 5 dengan Shannon *entropy*. (2) Membandingkan nilai rata-rata koefisien sinyal uji dan kata yang ingin diidentifikasikan dengan Euclidian *distance*.	(1) Fitur yang diekstraksi dari *wavelet transform* mempunyai akurasi paling tinggi untuk pengenalan kata. (2) Dekomposisi dengan *wavelet* Daubechies menghasilkan persentase tertinggi untuk pengenalan suara berbahasa Kannada.
4	Rui Ma (2014)	*Parametric Speech Emotion Recognition Using Neural Network*	Suara manusia dalam emosi senang, sedih, marah, dan netral.	(1) Normalisasi data suara. (2) Ekstraksi fitur berupa kecepatan berbicara, energi, *pitch*, frekuensi *formant*, dan MFCC. (3) Klasifikasi dengan *neural network*.	Akurasi klasifikasi dengan data baru mencapai 82.5%.

5	D.D. Joshi & M.B. Zalte (2013)	*Recognition of Emotion from Marathi Speech Using MFCC and DWT Algorithms*	Suara manusia berbahasa Marathi dalam emosi senang, sedih, marah, netral, takut, dan bosan.	(1) Evaluasi sampel suara oleh 10 pendengar. (2) Ekstraksi fitur suara dengan MFCC dan DWT db4 tingkat 4. (3) Klasifikasi dengan SVM.	Persentase pengenalan emosi masing-masing : marah (75%), bosan (84%), takut ($\pm$50%), senang (67%), sedih (89%), dan netral (64%).

Penggunaan *wavelet transform* dengan keluarga *wavelet* Daubechies untuk mengekstraksi fitur pada suara menghasilkan akurasi yang tinggi, baik untuk pengenalan kata (Anusuya & Katti, 2011) maupun pengenalan emosi melalui suara (Joshi & Zalte, 2013). Oleh karena itu, *wavelet transform* dengan keluarga *wavelet* Daubechies digunakan dalam penelitian ini. *Classifier neural network* dan SVM juga menunjukkan performa klasifikasi yang baik pada penelitian terdahulu, sehingga dimanfaatkan sebagai metode klasifikasi pada penelitian ini.

Database yang digunakan pada penelitian sebelumnya dibangun dengan menggunakan ucapan aktor dan aktris dari 60 film berbeda (Hamidi & Mansoorizade, 2012). Meskipun diasumsikan emosi pada film lebih alami, tidak menutup kemungkinan bahwa emosi diekspresikan secara berlebihan oleh aktor atau aktris tersebut. Selain itu, pada penelitian Ma (2014), *database* suara emosional dibangun dari suara peneliti sendiri sehingga karakteristik suara masing-masing emosi relatif sama. Agar emosi yang diekspresikan melalui vokal lebih alami, pada penelitian ini, responden diminta menonton klip film dan menceritakan pengalaman pribadi berkaitan dengan emosi yang ingin dipicu. Untuk menentukan fitur yang dapat digunakan dalam proses klasifikasi, penulis memanfaatkan *linear mixed effect models* sehingga hanya fitur yang dapat membedakan emosi yang digunakan untuk membentuk vektor masukan. Hal tersebut diharapkan dapat mengeliminasi fitur yang tidak dibutuhkan dan meningkatkan performa dari *classifier*.