Pengenalan Ucapan Kontinu Kosakata Besar Bahasa Indonesia Multimodal Berbasis Silabel Menggunakan Hidden Markov Toolkit dan Pergerakan Bibir

Hilman Fauzi Rijal[1], Dr. Suyanto, S.T., M.Sc.[2]

[1,2]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom, Bandung

[1] **hilmanfauzir40@gmail.com**, [2] **suyanto2008@gmail.com**

Abstrak

Pada pengenalan ucapan yang lumrah diaplikasikan adalah pengenalan ucapan terhadap sinyal suara. Namun, pada jenis data berupa video, data yang tersedia untuk diolah tidak hanya berupa sinyal suara saja. Maka, dibangunlah sistem Pengenalan Ucapan Kontinu Kosakata Besar berbasis silabel dengan menggabungkan fitur audio dan visual. Pembangunan sistem dilakukan dengan Hidden Markov Toolkit(HTK) dengan pengambilan fitur visual dengan menggunakan Discrete Cosine Transform(DCT) dan Principal Component Analysis(PCA). Hasil yang diperoleh dari sistem yang dibangun dapat memperkecil word error rate sebesar 6,07%.

Kata Kunci : Fitur Visual, HTK, DCT, PCA, Pengenalan Ucapan Kontinu, Silabel

Abstract

The common thing applied on speech recognition is speech recognition on the speech signal. However, on the type of data in the form of video, the data which is available is not only a speech signal. Therefore, Syllable Based Large Vocabulary Continous Speech system with fusion of audio and visual feature is developed. The system is developed with the use of Hidden Markov Toolkit(HTK) along with visual feature extraction using Discrete Cosine Transform(DCT) and Principal Component Analysis(PCA). The result obtained from the system is able to reducing word error rate 6,07%.

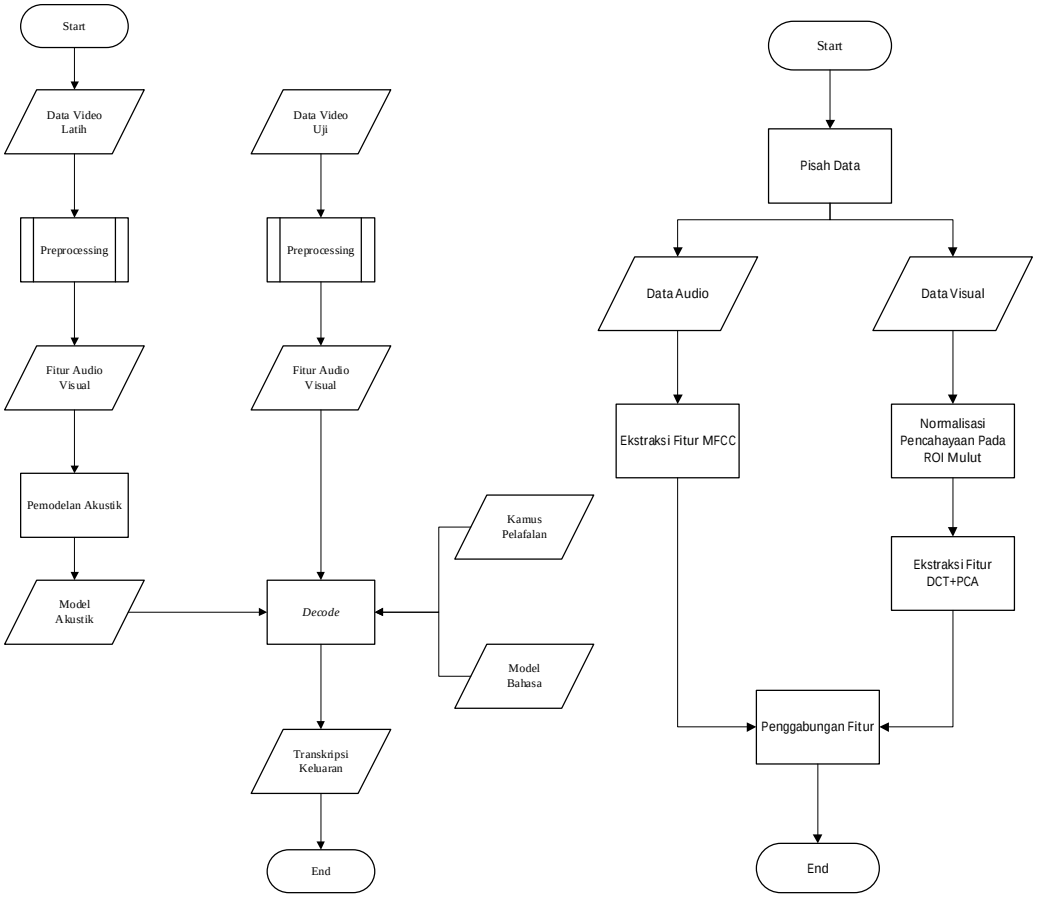
Keywords : Visual Feature, HTK, DCT, PCA, Continous Speech Recognition, Syllable

1. Pendahuluan

Penggunaan video sebagai media semakin marak. Banyaknya platform yang menyediakan layanan berbagi video menyebabkan siapapun dapat mengunggah video di Internet. Salah satu platform berbagi video terbesar adalah Youtube. Youtube memiliki miliaran pengguna, hampir sepertiga dari pengguna internet dan setiap harinya pengguna melihat ratusan juta jam video [1]. Bahkan sekitar 500 jam video yang diunggah pada Youtube setiap menitnya pada tahun 2015 [2]. Pada tahun 2015, 70% dari pengguna internet mengakses video dan pada tahun 2020 diprediksikan akan mencapai 82% [3]. Pada tahun 2010, Janakiraman dkk. menunjukan pengenalan ucapan berdasarkan silabel mampu menurunkan kompleksitas perhitungan dan word error rate(WER) menjadi 4,4% pada database TIMIT dan 21,2% pada database NTIMIT [4]. Pada data video, akan banyak ditemukan pengucapan kontinu. Sehingga untuk pengolahan ucapan pada data video, dengan menggunakan sinyal ucapannya saja, maka masih tersisa data yang dapat dimanfaatkan. Maka dibangun sistem Pengenalan Ucapan Kontinu Berbasis Silabel dengan pengambilan fitur pada gerakan bibir. HTK dikembangkan oleh Cambridge University Engineering Department (CUED). HTK ditujukan untuk membangun dan memanipulasi Hidden Markov Model (HMM). Kegunaan utama HTK untuk penelitian pengenalan ucapan meskipun dapat digunakan pada apapun yang menggunakan data time series [5]. HTK digunakan untuk membangun sistem pengenalan ucapan.

2. Perancangan Sistem

Pada gambar 1, digambarkan secara umum sistem yang dibangun pada penelitian ini.Pemodelan akustik, model bahasa, dan decode dilakukan menggunakan HTK.



Gambar 1 Flowchart Sistem Pengenalan Ucapan Multimodal Audio Visual

Pada preprocessing, ekstraksi fitur dilakukan pada data visual maupun suara. Pada data visual, proses ekstraksi fitur dilakukan pada tiap frame. Pada tahap awal, dilakukan deteksi facial landmark menggunakan Dlib untuk mendapatkan ROI pada mulut. Dari ROI mulut tersebut, dilakukan resize sehingga berdimensi 64x64. Untuk menormalisasi cahaya pada gambar, pada penelitian[6] Wang dkk berhasil melakukan pengenalan wajah pada dataset Yale dengan akurasi 98,3% dengan metode *weberface*. Kemudian *weberface* dipilih untuk diaplikasikan pada ROI mulut untuk normalisasi cahaya untuk kemudian dilakukan ekstraksi fitur menggunakan DCT.

Untuk ekstraksi fitur menggunakan DCT, digunakan matrix DCT berdimensi 8x8. Pada kompresi standar JPEG, kompresi berdasarkan matriks 8x8 menghasilkan gambar dengan kualitas terbaik [7]. Pada ROI mulut, untuk mendapatkan koefisien DCT, gambar dibagi oleh block 8x8 menjadi sebanyak 64 block pada gambar 64x64. Pada setiap block, diambil koefisien AC pada DCT dengan frekuensi yang rendah.

3. Skenario Pengujian

Pengujian dilakukan untuk mendapatkan performansi terbaik dari sistem yang telah dibangun. Dilakukan Pengujian terhadap jumlah koefisien AC yang akan diambil pada fitur visual. Jumlah koefisien AC yang diambil adalah 2, 5, 9, 14, 20, 27, dan 35.

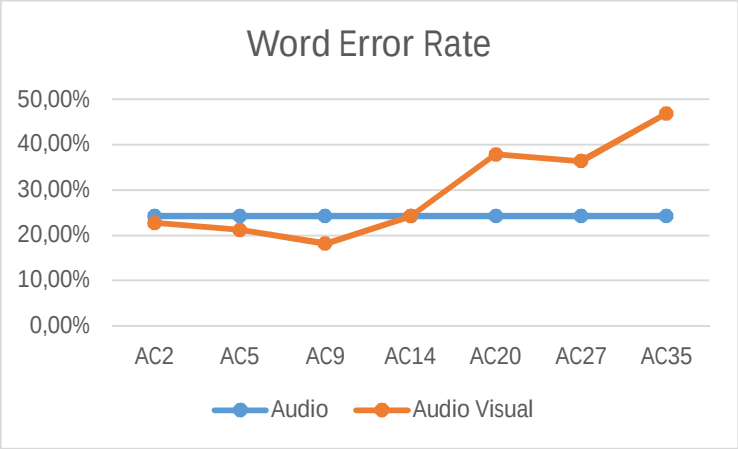
4. Hasil dan Analisis

Pada sistem pengenalan ucapan audio dan multimodal audio-visual, dilakukan evaluasi *Word error Rate (WER)* data audio dan audio visual dengan masing – masing jumlah koefisien AC yang diujikan pada tabel 1.

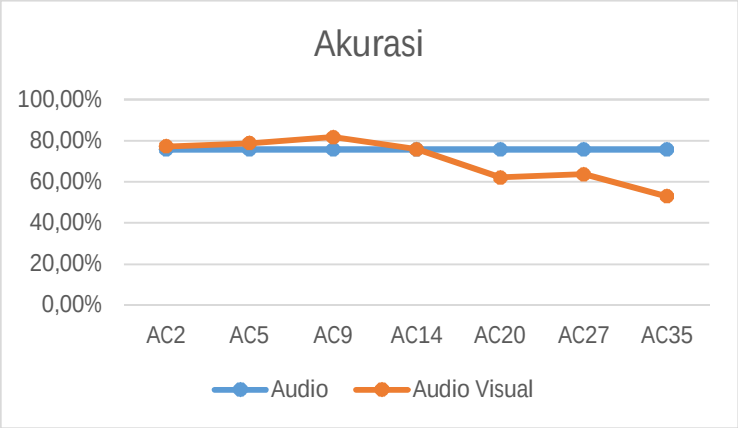
Jenis Data / Jenis Error	Audio	AC2	AC5	AC9	AC14	AC20	AC27	AC35
Insertions	9,09%	10,6%	9,09%	9,09%	13,63%	25,75%	21,21%	33,3%
Deletions	12,12%	7,57%	6,06%	4,54%	3,03%	1,51%	3,03%	0%
Subtitutions	3,03%	4,54%	6,06%	4,54%	7,57%	10,6%	12,12%	13,63%

Tabel 1 Tabel WER pada Data Audio dan Audio Visual dengan jumlah koefisien AC yang digunakan

Jika dibuatkan grafik, maka grafik perbandingan WER antara data audio dan audio visual pada sistem ini dapat dilihat pada gambar 2. Dan untuk perbandingan akurasi pada gambar 3.



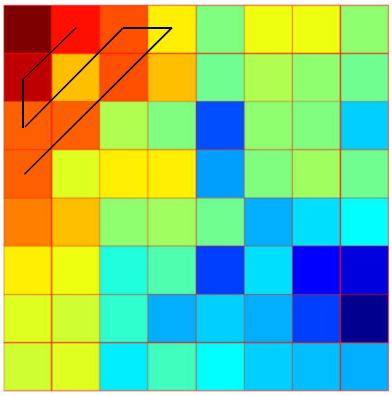
Gambar 2 Grafik Perbandingan WER antara fitur Audio dan Audio Visual



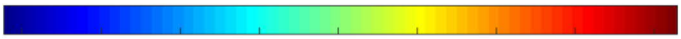
Gambar 3 Grafik Perbandingan Akurasi antara fitur Audio dan Audio Visual

Penggunaan koefisien AC pada data audio visual sebanyak 9 dapat dilihat pada gambar 4-2 mempunyai WER paling rendah dengan nilai 18,17% dibandingkan dengan data audio yang bernilai 24,24%. Dapat diamati juga untuk penggunaan koefisien AC selain 9 menghasilkan nilai lebih rendah meskipun penggunaan koefisien AC 2, 5, dan 14 masih memiliki nilai lebih baik daripada data audio yaitu, 22,71%, 21,21%, dan 24,23%. Sedangkan untuk akurasi, penggunaan koefisien AC 9 mempunyai akurasi 81,82% dibandingkan dengan data audio mempunyai akurasi sebesar 75,76%.

Berdasarkan hasil pengujian, penggunaan koefisien AC 9 lebih unggul dari penggunaan koefisien AC yang lain. Meskipun koefisien AC 9 mengambil koefisien AC dengan frekuensi lebih tinggi dari penggunaan koefisien AC 2 dan 4. Diambil block koefisien DCT pada proses ekstraksi fitur yang dapat dilihat pada gambar 4 dengan *colormap* pada gambar 5 di mana biru merepresentasikan koefisien terendah dan merah tertinggi, dapat diamati pada diagonal terakhir masih terdapat koefisien AC yang tinggi. Sedangkan untuk pengambilan koefisien AC pada diagonal seterusnya, nilai koefisien cenderung semakin menurun.



Gambar 4 Koefisien DCT pada block matriks 8x8



Gambar 4-1 *Colormap* koefisien DCT

5. Kesimpulan

Berdasarkan hasil pengujian, kesimpulan yang didapat pada tugas akhir ini adalah:

1. Dengan penambahan fitur visual untuk melakukan pengenalan ucapan kontinu berhasil menurunkan WER sebesar 6,07%.
2. Jumlah penggunaan koefisien AC perlu diperhatikan, sehingga koefisien bernilai tinggi yang digunakan sebagai fitur.

Daftar Pustaka:

[1] Youtube, *YouTube for Press*, [Online] Available At: https://www.youtube.com/yt/press/en-GB/statistics.html [Accessed 22 March 2017]

[2] Robertson, M.R., 2015, *500 Hours of Video Uploaded to Youtube Every Minute [Forecast]*, [Online] Available At: http://tubularinsights.com/hours-minute-uploaded-youtube/ [Accessed 22 March 2017]

[3] Cisco, 2017, *Cisco Visual Networking Index: Forecast and Methodology, 2016–2021*, White paper Cisco public

[4] Janakiraman, R., Kumar, J. C., Murthy, H. A., 2010, *Robust syllable segmentation and its application to syllable-centric continuous speech recognition*, Chennai, National Conference On Communications (NCC).

[5] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D. and Valtchev, V., 2002. *The HTK book*. Cambridge university engineering department, 3, p.175.

[6] Wang, B., Li, W., Yang, W., Liao, Q., 2011, *Illumination Normalization Based on Weber's Law With Application to Face Recognition*, IEEE Signal Processing Letters, vol. 18, no. 8, pp. 462-465.

[7] Wallace, G.K.,1992, *The JPEG Still Picture Compression Standard*, Massachusetts, IEEE Transactions on Consumer Electronics, Vol. 38, No. 1.