Nama : Suryadi
NPM : 1706071245

# Rangkuman

## Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models

Paper ini menjelaskan mengenai penerapan spontaneous speech recognition menggunakan speech corpus Bahasa Indonesia. Beberapa kesulitan yang muncul ketika menggunakan spontaneous speech recognition dikarenakan spontaneous speech mengandung penundaan kata, penghubung kata, kesalahan penyebutan kata, dan urutan kata yang berubah-rubah.

Spontaneous speech recognition menggunakan speech corpus dengan banyak pengejaan masing-masing kata dengan variasi logat yang berbeda-beda. Adaptasi dari akustik model ini bertujuan untuk meningkatkan akurasi spontaneous speech recognition. Teknik yang digunakan untuk adaptasi akustik model ini berdasarkan Gaussian Mixture and Hidden Markov Models (GMM-HMM) yaitu

1. Maximum A-posteriori Probability (MAP)
2. Maximum Mutual Information (MMI)
3. Feature-space Maximum Likelihood Linear Regression (fMLLR)

MAP dan MMI sangat berperan dalam mengurangi word error rate untuk spontaneous speech recognition.

Speech Corpus yang digunakan pada implementasi machine learning terdiri dari :

1. Dictated speech training set
2. Spontaneous speech training set
3. Dictated speech evaluation set
4. Spontaneous speech evaluation set

Dengan rangkuman data speech corpus sebagai berikut

Table 2. Summary of the speech corpora.

| Corpus | #Speakers | #Sentences | Duration |
|---|---|---|---|
| READ-TR | 244 | 63,291 | 72h55m50s |
| SPONTAN-TR | 11 | 1,058 | 43m35s |
| UNION-TR | 255 | 64,349 | 73h39m25s |
| READ-EV | 20 | 6,682 | 14h31m45s |
| SPONTAN-EV | 28 | 1,085 | 48m38s |

Nama : Suryadi
NPM : 1706071245

Langkah-langkah yang dilakukan sebelum melakukan implementasi spontaneous speech recognition

1. Building Dictionary

Pembuatan kamus bahasa dengan penyesuaian masalah pada spontaneous speech recognition seperti alternatif pengucapan, standarisasi pengucapan, memperbaiki kesalahan pengucapan,dll.

Table 3. List of all phonemes utilized in the HMM training.

| Our symbol | IPA symbol | Our symbol | IPA symbol | Our symbol | IPA symbol |
|---|---|---|---|---|---|
| a | /a/ | i | /i/ | q | /q/ |
| ai | /aɪ̯/ | j | /dʒ/ | r | /r/ |
| au | /aʊ̯/ | k | /k/ | s | /s/ |
| b | /b/ | kh | /x/ | sy | /ʃ/ |
| c | /tʃ/ | l | /l/ | t | /t/ |
| d | /d/ | m | /m/ | u | /u/ |
| e | /e/ | n | /n/ | w | /w/ |
| @ | /ə/ | ng | /ŋ/ | y | /y/ |
| f | /f/ | ny | /ɲ/ | z | /z/ |
| | /v/ | o | /o/ | SIL | |
| g | /g/ | oi | /oɪ̯/ | SPN | |
| h | /h/ | p | /p/ | NSN | |

2. Building Language Model

Pembuatan Language Model menggunakan evaluasi akustik model dari pelatihan machine learning. Pelatihan machine learning menggunakan penggabungan data dari 2 speech corpus (dictated dan spontaneous ) dengan rasio tertentu.

Table 4. Best mixing weight for both scenarios.

| Scenario | READ-LM | SPONTAN-LM | Total |
|---|---|---|---|
| Dictated | 0.994006 | 0.005994 | 1 |
| Spontaneous | 0.542802 | 0.457198 | 1 |
| **Mean** | 0.768404 | 0.231596 | 1 |

Table 5. Statistics of the training and evaluation text.

| Field | Tala 2003 | SPONTAN-TR | READ-EV | SPONTAN-EV |
|---|---|---|---|---|
| #Unique Sentences | 613,054 | 1,052 | 343 | 1,085 |
| #Words | 10,250,367 | 5,265 | 4,016 | 5,684 |
| OOV Rate | | | 4.18% | 0.58% |

Table 6. Perplexities for both scenarios.

| LM | Dictated ppl | Spontaneous ppl |
|---|---|---|
| READ-LM | 219.76 | 887.35 |
| SPONTAN-LM | 18,038.90 | 1,325.39 |
| INTER-LM | 240.54 | 469.30 |

Nama : Suryadi
NPM : 1706071245

3. Building Acoustic Model

Pembuatan akustik model berdasarkan Gaussian Mixture and Hidden Markov Models (GMM-HMM) dilakukan untuk feature extraction dengan MAP, MMI, fMMLR training.

Hasil implementasi spontaneous speech recognition ditampilkan dalam parameter Word Accuracy Rate (WAR).

$$WAR = \left(1 - \frac{S + I + D}{N}\right) \times 100\%$$

S = Substitution

I = Insertion

D = Deletion

N = Jumlah Kata pada test set

Data hasil experiment dapat dilihat pada table dibawah

Table 7. Experiment result in % WAR.

| Acoustic Model | SPONTAN-EV | READ-EV |
|---|---|---|
| Triphone | 54.26 | 81.53 |
| Triphone + MAP | **56.86** | 82.89 |
| Triphone + MMI | 53.80 | **83.01** |
| Triphone + MAP + fMLLR (SI) | 41.19 | 21.46 |
| Triphone + MAP + fMLLR (SD) | 56.79 | 56.68 |
| Triphone + MMI + fMLLR (SI) | 42.19 | 19.88 |
| Triphone + MMI + fMLLR (SD) | 56.74 | 56.68 |

Note: SI = speaker-independent model
SD = speaker-dependent model

Model triphone yang diadaptasi MAP paling akurat dalam skenario spontaneous, sedangkan model yang diadaptasi MMI berkinerja lebih baik dalam skenario dictated. Model akustik dibuat untuk mengenali ucapan dictated dan spontaneous. Adaptasi MAP menghasilkan peningkatan absolut 2,60% dan 1,36% dalam Word Accuracy Rate (WAR).

Nama : Suryadi
NPM : 1706071245

# Rangkuman

## Building a Speech and Text Corpus of Turkish : Large Corpus Collection with Initial Speech Recognition Results

Fungsi utama dari Automatic Speech Recognition (ASR) yaitu mengubah pengucapan manusia menjadi teks tertulis. Implementasi ASR dapat digunakan pada data social media, kalimat perintah untuk perangkat, Subtitle dari film, dll. ASR bekerja dengan model yang dihasilkan dari pelatihan machine learning berdasarkan klasifikasi pola statistik. ASR menggunakan supervised learning untuk melatih speech classifier. Pelatihan ASR ini menggunakan metode Gaussian Mixture Model – Hidden Markov Models (GMM-HMM). Meskipun metode ini sudah sukses diterapkan tetapi akurasi ASR masih terbatas untuk kemampuan manusia berbicara.

Sebelum melakukan implementasi ASR perlu dilakukan pengumpulan data speech corpus. Metode yang digunakan untuk mengumpulkan data speech corpus sebagai berikut :

1. Use of Film Movies and Time-Bound Subtitle Documents (subtitle film)
2. Collections of Speech Data with a Mobile Application (google translate text to speech)
3. Collections of Speech Data with Transfer Learning (data berasal dari machine learning lainnya)

Langkah-langkah implementasi ASR

1. Speech Corpus for Experiments

   Corpus mengandung banyak variasi suara dengan frekuensi sampling 16 kHz dan 16 bit sample size

**Table 3.** Information on Turkish corpora.

| Corpus Name | Duration (h) | Total Number of Utterances |
|---|---|---|
| METU | 8.33 | 8002 |
| Bogazici | 94.44 | 82,033 |
| Unverified HS | 460.12 | 780,014 |
| Verified HS | 350.27 | 565,073 |

2. Development of the Turkish ASR system for experiments

Toolkit Kaldi digunakan untuk pengembangan sistem ASR Turki. Kaldi merupakan toolkit open source digunakan untuk aplikasi ASR dan ditulis dalam Bahasa pemrograman C++. Model ASR menggunakan GMM-HMM dan Deep Neural Network (DNN).

Nama : Suryadi
NPM : 1706071245

Performa dari ASR ini dapat diukur menggunakan Word Error Rate (WER).

$$WER = \frac{D + S + I}{N} \times 100$$

S = Substitution

I = Insertion

D = Deletion

N = Jumlah Kata pada test set

Dari hasil experiment didapatkan data sebagai berikut

**Table 4.** Results of Gaussian mixture model (GMM) based Turkish automatic speech recognition (ASR) systems.

| Corpus Name | Word Error Rate (WER%) |
|---|---|
| METU | 70.71 |
| Bogazici | 27.70 |
| Unverified HS | 55.27 |
| Verified HS | 24.70 |

**Table 6.** Results of deep neural networks (DNN)-based Turkish ASR systems.

| Corpus Name | Word Error Rate (WER%) |
|---|---|
| METU | 64.55 |
| Bogazici | 22.63 |
| Unverified HS | 49.20 |
| Verified HS | 18.70 |

Dari kedua data diatas dapat disimpulkan bahwa model akustik DNN memberikan performa yang lebih baik dibandingkan GMM-HMM karena WER yang semakin kecil.