MENGAPA MESIN PENCARI SUARA GAGAL MENGENALI BAHASA INDONESIA? SEBUAH KAJIANAWAL TENTANG ASR (*AUTOMATIC SPEECH RECOGNITION*) BAHASA INDONESIA

Oleh,
Totok Suhardiyanto
Universitas Indonesia
Email: totok.suhardiyanto@ui.ac.id

Abstract

This paper is about the study of Indonesian Automatic Speech Recognition (ASR) designed by Informational- Technological Computer (TIK).Specifically, this paper is aimed at describing how this tool operates in recognizing some in-puts in Indonesian language. TIK industry has something to do with Indonesian PWO where the use of the smart phones developes massively in Indonesia. This study processes around 10.774 data in Indonesian language in form of sentence, phrase, and word. From this number, less than 20% can be categorized perfect. The others have an error in format and recognition. This is due to some factors bringing about the failure of ASR in recognising the in put in Indonesian language.

Key words: ASR (Automatic Speech Recognition), voice search, format error, recognation failure, WER (Word Error Rate)

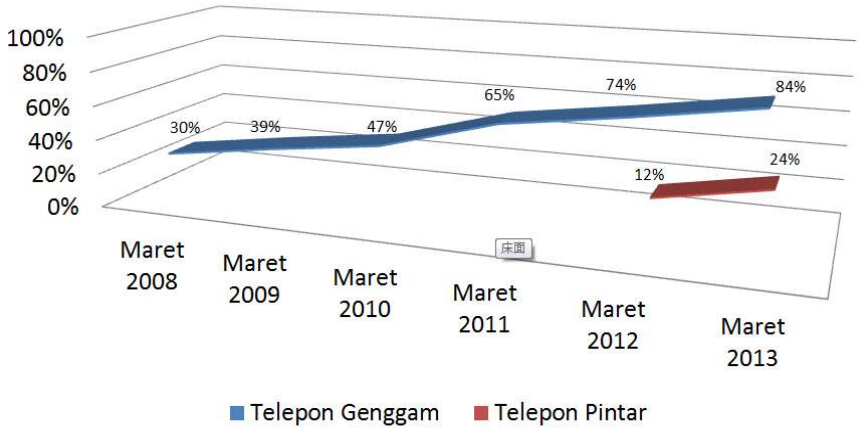
1. Pendahuluan

Dalam kajian linguistik komputasional, pengenalan wicara (*speech recognition*) merupakan salah satu bidang yang sangat menantang, khususnya bagi bahasa-bahasa non-Eropa. Bidang ini merupakan salah satu kajian utama dalam bidang linguistik komputasional dan bersentuhan dengan intelijensi buatan (*artificial intelligence*) (Jurafsky and Martin, 2008). Pengenalan wicara pada umumnya dapat dibatasi sebagai sistem yang mampu mengubah masukan dalam bentuk suara ke dalam bentuk tertulis, atau mentranskripsikan bunyi ke bentuk tulisan. Selanjutnya, pengenalan wicara dalam makalah ini akan dirujuk sebagai ASR (*Automatic Speech Recognition*), singkatan yang lazim digunakan di kalangan peneliti linguistik komputasional dan pemrosesan bahasa alami (NLP = *natural language processing*).

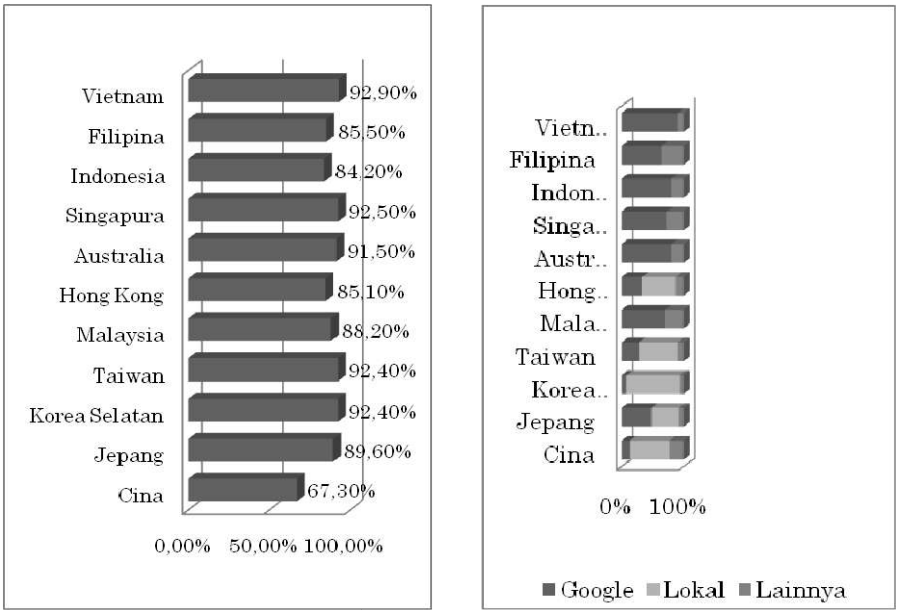
Setakat ini, dapat dikatakan bahwa perkembangan ASR Bahasa Indonesia masih belum menggembirakan. Kualitasnya masih cukup jauh dari harapan. Meskipun demikian, beberapa nama besar dalam industri TIK sangat menaruh perhatian dalam persoalan ASR Bahasa Indonesia (selanjutnya disingkat ASR BI). Goodwater dkk. (2010)

mengatakan bahwa tuturan merupakan salah satu hal tersulit untuk dikenali dalam kajian sistem pengenalan otomatis karena tingginya tingkat ketidaklancaran, variasi pelafalan, sertabanyaknya variabel dalam hal akustik dan prosodi. Lebih lnjut, Goodwater dkk menyebut ada tiga faktor yang potensial berpengaruh pada tingkat kesalahan ASR, yaitu faktor ketidaklancaran, faktor leksikal, dan faktor prosodis. Oleh karena itu, sangat penting untuk mengetahui yang mana dari faktor-faktor tersebut yang bermasalah bagi ASR BI.

Mengapa kajian ASR BI menjadi sangat penting dan menjadi salah satu perhatian industri TIK internasional? Ada dua alasan utama. Pertama, ke depan hampir semua peranti TIK, terutama telepon pintar (*smartphone*) akan menggunakan sistem berbasis suara (*voice generated system*). Jadi, basis interaksi di antara pengguna manusia dan mesin TIK tidak lagi berupa masukan berbentuk huruf atau *string*, melainkan suara.

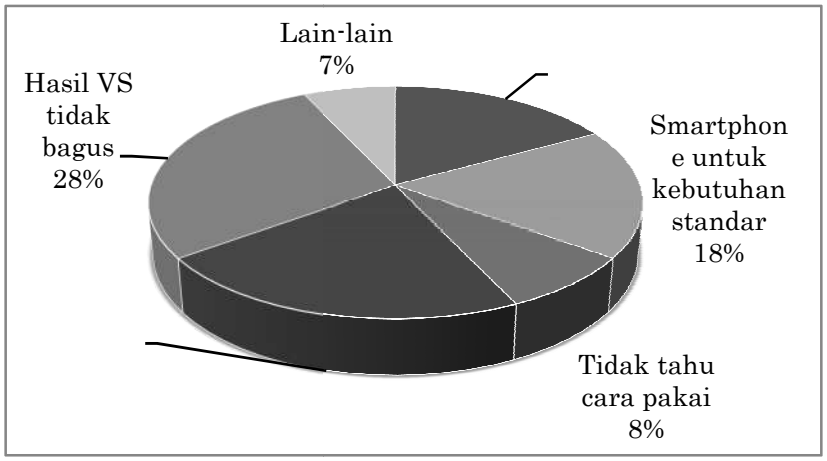


Gambar 1: Kepemilikan telepon genggam dan telepon pintar di Indonesia 2008- 2013 (sumber: http://www.roymorgan.com/findings)



Gambar 2: Perbandingan jumlah pengguna mesin pencari di wilayah Asia Timur Sumber http://www.comscore.com/analysis)

Aplikasi ASR memang belum begitu populer di Indonesia. Berdasarkan survei yang dilakukan dalam penelitian ini, paling tidak terdapat empat alasan utama mengapa di Indonesia fenomena kenaikan jumlah pengguna telepon pintar secara tajam tidak berbanding lurus dengan meningkatnya pemakaian VS?



Gambar 3: Hasil survei mengenai alasan tidak menggunakan VS pada telepon genggam (N= 33)

Alasan pertama adalah kampanye atau publikasi mengenai penggunaan mesin pencari suara tidak optimal. Tidak banyak orang yang tahu bahwa beberapa perusahaan TIK, termasuk Google dan Microsoft, telah menyediakan aplikasi berbasis suara (pada Gambar 3: "Tidak tahu ada VS"). Kedua, hasil dari mesin pencari suara bahasa Indonesia tidak bagus dan tidak bisa diandalkan ("Hasil VS tidak bagus").. Akibatnya, banyak pengguna telepon pintar di Indonesia yang tidak mau lagi menggunakannya. Ketiga, dalam menggunakan telepon genggam atau telepon pintar, orang Indonesia lebih suka memasukkan input dengan cara mengetikkannya daripada mengucapkannya ("Pencarian teks lebih nyaman"). Barangkali ini terkait dengan keutamaan menggunakan pesan singkat secara tertulis daripada pesan dengan suara di Indonesia. Pesan singkat masih dapat digunakan dan ditolerir pada saat rapat, namun tidak untuk pesan bersuara. Keempat. pada dasarnya masyarakat Indonesia masih belum dapat dikategorikan masyarakat yang berteknologi tinggi. Mereka mampu membeli peranti yang lebih canggih, seperti telepon pintar misalnya, namun sebagian besar hanya menggunakan fungsi-fungsi dasar dari peranti generasi sebelumnya ("Menggunakan smartphone untuk kebutuhan standar").

Secara umum, pengguna telepon pintar beranggapan bahwa ASR BI masih jauh dari kondisi yang memuaskan (lihat Gambar 3). Paling tidak hal tersebut terkait dengan beberapa faktor baik teknis maupun linguistis. Pertama adalah sistem aplikasi yang belum terbangun dengan sempurna. Kaidah algoritma, pemodelan, maupun pangkalan data masih belum cukup untuk mengenali masukan berbahasa Indonesia. Kedua, sistem jaringan internet 3G di Indonesia masih belum ajek dan mempunyai kapasitas merata di seluruh wilayah Indonesia. Akibatnya, penangkapan sinyal audio dan transfer data sangat rentan mengalami gangguan. Ketiga, belum ada variasi lafal yang menjadi standar dalam kasus bahasa Indonesia. Sebagai akibatnya, perancang ASR mengembangkan ASR BI tidak berdasarkan data variasi bahasa di lapangan.

Selain pemain internasional, sistem pengenalan suara bahasa Indonesia atau ASR BI juga pernah dikembangkan oleh peneliti lokal (Suyanto & Adityatama 2012). Suyanto dan Adityatama membangun sistem aplikasi yang mereka namai IndoVM untuk mendiktekan pesan singkat berbahasa Indonesia pada telepon pintar Android. Penelitian mereka masih menunjukkan tingginya tingkat kesalahan kata atau word error rate (*WER*) untuk sistem yang berbasis model statistik (*statistical language model*) dan cukup rendah pada sistem yang berbasis kaidah bahasa. Aplikasi tersebut menampilkan performa yang cukup bagus baik dari segi akurasi, maupun waktu tanggapan.

Kajian terdahulu pada pengenalan monolog dan dialog spontan, serta

percakapan menunjukkan bahwa kata-kata yang jarang digunakan cukup sulit dikenali (Fosler-Lussier & Morgan 1999; Shinozaki & Furui 2001). Selain itu, tingkat kesalahan yang tinggi juga terkait dengan tuturan atau ujaran yang cepat (Siegler & Stern 1995; Fosler-Lussier & Morgan 1999; Shinozaki & Furui 2001).

Dalam kajian Shinozaki & Furui (2001) terhadap ASR bahasa Jepang, tuturan yang sangat lambat dan panjang kata (dalam bentuk suara) juga berpengaruh pada tingkat kesalahan pengenalan. Kata-kata yang lebih pendek ternyata lebih sering mengalami kesalahan. Kemudian, Adda-Decker & Lamel (2005) juga mengemukakan bahwa baik ASR bahasa Perancis maupun Inggris menemui kesulitan dalam mengenali tuturan laki-laki daripada perempuan karena terkait dengan tingkat ketidaklancaran dan penggunaan kata-kata yang dipendekkan.

Sejalan dengan penelitian pada ASR, hasil penelitian dalam bidang psikolinguistik pun menyebutkan bahwa seperti ASR, manusia pun lebih mudah mengenali—tentu saja lebih cepat dan lebih akurat—kata yang lebih sering digunakan daripada kata yang jarang digunakan (Marslen-Wilson 1987; Dahan dkk. 2001). Selain itu, pengenalan juga lebih sulit pada kata yang secara fonetis mirip dengan kata lain daripada kata yang mempunyai tingkat distingtif tinggi (Luce & Pisoni 1998).

Dalam ASR BI, tentu saja juga menarik apakah faktor-faktor yang ditemukan dalam penelitian terdahulu juga muncul secara signifikan. Oleh sebab itu, kertas kerja ini disusun atas dasar dua pertanyaan mendasar. Pertama, bagaimana kinerja atau performa ASR BI dalam proses pengenalan masukan berbahasa Indonesia. Kedua, faktor-faktor apa saja dalam bahasa Indonesia yang berpengaruh terhadap kinerja atau performa tersebut.

2. Metodologi

2.1 Data

Pada kertas kerja ini,analisis didasarkan pada keluaran (*output*) mesin pencari suara atau VS dari sebuah perusahaan TIK terkemuka. Diperoleh data sebanyak 10.774 unit pencarian yang terdiri atas kata, frase, klausa, dan bahkan kalimat. Pemerolehan data tersebut dilakukan secara acak dalam kurun waktu sebulan pada Maret 2013.

Untuk menjawab pertanyaan pertama pada penelitian ini, yakni bagaimana kinerja atau performa ASR tersebut, pertama kali dilakukan klasifikasi kesalahan ke dalam tigakategori utama: kategori salah, kategori benar, dan kategori cacat data. Kategori salah merupakan keluaran ASR BI yang tidak sesuai dengan rujukan atau masukan. Kategori ini dibagi lagi menjadi tiga subkategori yakni pengurangan, penambahan, dan penyulihan (lihat Gambar 4).

```
REF:    tetapi DIA tidak mau  membeli    ***   kamus
ASR:    tetapi BILA tidak    ***membeli SEBUAH kamus
Eval:          S          D          I
```

Gambar 4: Contoh kategorisasi data ke dalam penambahan (I), penyulihan (S), dan pengurangan (REF= referensi atau masukan, ASR= hasil pengenalan atau keluaran, dan Eval= evaluasi dan kategorisasi).

Sementara itu, kategori benar adalah keluaran yang dihasilkan ASR tepat atau sesuai dengan rujukan masukan. Kategori terakhir,yakni kategori cacat data, adalah data yang tidak dapat digunakan karena kualitas audio yang buruk.

Setelah dilakukan klasifikasi data, kategori benar dan cacat data dikeluarkan dari analisis karena fokus penelitian ini pada kesalahan pengenalan ASR BI. Meskipun demikian, jumlah keluaran benar yang dihasilkan ASR digunakan lagi ketika menghitung F measure untuk membandingkan hasilnya dengan penghitungan tingkat kesalahan kata atau WER.

Untuk menjawab pertanyaan kedua, dalam analisis digunakan standar pengukuran *word error rate* (SER) dengan formulasi sebagai berikut.

$$\mathbf{WER} = \frac{s + d + i}{REF}$$

dengan perincian
s = penyulihan
d = penyulihan
i = penambahan/penggantian
REF = jumlah total data yang menjadi referensi

2.2 Analisis Fitur

Pada bagian ini, dijabarkan fitur yang akan dianalisis pada data kesalahan yang telah diperoleh dan dikategorisasikan. Pada makalah ini, hanya hasil analisis dua fitur yang akan ditampilkan. Kedua fitur itu adalah ketidaklancaran dan kategori kata.

Sebagaimana disebutkan sebelumnya, ketidaklancaran dipercayai dapat meningkatkan kesalahan pengenalan oleh sistem ASR (Goldwater 2010). Fitur ketidaklancaran yang diperhatikan dalam penelitian ini terdiri atas

(1) *sebelum/setelah jeda*
Fitur ini merujuk pada kata yang muncul langsung mendahului atau mengikuti sebuah jeda
(2) *sebelum/setelah fragmen*
Fitur ini merujuk pada kata yang muncul langsung mendahului atau mengikuti sebuah fragmen
(3) *sebelum/setelah repetisi*

Fitur ini merujuk pada kata yang muncul langsung mendahului atau mengikuti sebuah pengulangan

(4) *dalam repetisi*

Fitur ini merujuk pada kata yang muncul dalam rangkaian pengulangan.

Keluaran ASR	**Fitur**
ya	sebelum pengulangan
anak	repetisi pertama
anak	repetisi kedua
anak	repetisi ketiga
dari	setelah repetisi
kebon	sebelum jeda
mmm	
kacang	setelah jeda
tanah	sebelum fragmen
bub	
abang	setelah fragmen

Karena keluaran ASR yang dijadikan data dalam penelitian ada yang berbentuk kata tunggal, frase, klausa, bahkan kalimat, analisis fitur ketidaklancaran ini hanya akan diterapkan pada data yang tidak berbentuk kata tunggal.

Selanjutnya, beberapa penelitian terdahulu juga menyebutkan bahwa kata yang jarang muncul lebih sulit untuk dikenali. Selain itu, ada juga yang mengemukakan bahwa kata individual dan kelompok kata juga berbeda dalam hal pengaruhnya terhadap kesalahan ASR. Oleh karena itu, dalam penelitian ini, fitur kategori leksikal juga akan diamati. Fitur ini terdiri atas dua kelompok.

Kelompok pertama adalah fitur sintaksis yang terdiri atas tiga kelas, yaitu:

(1) kelas terbuka,
(2) kelas tertutup, dan
(3) kelas pemarkah wacana.

Nomina, verba, dan adjektiva masuk ke dalam kelas terbuka, sementara preposisi, artikula, konjungsi, dan sebagainya masuk ke dalam kelas tertutup. Kelas Pemarkah wacana mencakup unsur-unsur seperti: *wah*, *deh*, *kok*, dan sebagainya.

Kelompok kedua adalah fitur lingkungan kata yang terdiri atas dua hal, yakni

(1) kata individual, dan
(2) rangkaian kata.

Fitur kata individual merupakan kata kunci pencarian yang hanya terdiri atas satu kata tunggal. Sementara itu, fitur rangkaian kata merupakan fitur yang mencakupbentuk yang bukan merupakan kata tunggal, melainkan bentuk yang muncul bersama dengan unsur lain.

3. ASR Bahasa Indonesia

Sebelum masuk ke dalam bagian hasil dan pembahasan, ada baiknya pada makalah ini dijelaskan apa dan bagaimana ASR. Secara ringkas, ASR atau pengenalan wicara otomatis merupakan sistem atau aplikasi yang berkemampuan untuk menerjemahkan bahasa lisan ke dalam bentuk tulisan (Jurafsky and Martin 2008).

ASR masa kini pada umumnya dikembangkan berdasarkan model statistik agar dapat menghadapi variasi dalam dialek, aksen, suara latar, dan pelafalan. Jika dijalankan pada lingkungan yang sunyi, ASR dengan model ini dapat mencapai ketepatan lebih dari 90%. Meskipun tiap ASR yang dikembangkan perusahaan IT berbeda satu sama lain, namun paling tidak ada model yang biasanya digunakan untuk memrosesmasukan wicara di dalam ASR.

(1) Berbasis contoh
Model ini menggunakan pangkalan data yang disematkan ke dalam program. Setelah memasukan input suara ke dalam sistem, pengenalan berlagsung dengan mencocokkan input ke pangkalan data. Kekurangan model ini terletak pada ketidakmampuan sistem untuk mengenali input yang tidak ada di dalam pangkalan data.

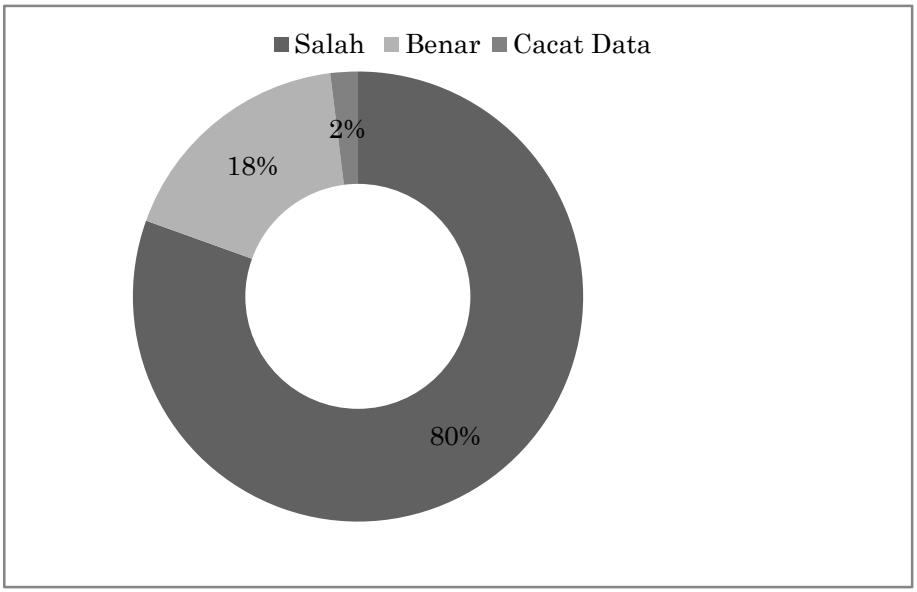
(2) Berbasis pengetahuan
Model ini menganalisis spektrogram wicara untuk memperoleh data dan menciptakan kaidah yang mampu mengatasi input atau masukan wicara yang diterima ASR.

(3) Stokastik
Model ini merupakan yang paling umum pada saat ini. ASR stokastik menggunakan model probabilitas untuk memodelkan ketidakapastian input.

ASR BI yang menjadi sumber data dalam penelitian ini merupakan ASR yang dikembangkan dengan model berbasis pengetahuan.

4. Hasil dan Diskusi

Hasil klasifikasi terhadap data dalam penelitian ini menunjukkan bahwa kesalahan masih sangat mendominasi hasil pengenalan ASR terhadap masukan berbahasa Indonesia. Dari 10.774 kata, hanya 17.55% kata yang dikenali dan sesuai dengan masukan, sementara 80.46% salah dan tidak sesuai dengan masukan. Hanya sekitar 1.98% data yang masuk kategori cacat data (lihat Gambar 5).



Gambar 5: Hasil klasifikasi data keluaran ASR BI

Dalam mengukur performa ASR BI, setelah diperoleh nilai WER, untuk memberikan perbandingan dalam penyajian pada Tabel 1 diberikan pula nilai F measure yang biasanya digunakan untuk mengukur keakuratan mesin pencari. Jadi, jika WER menunjukkan kesalahan, F measure justru sebaliknya.

Dari sudut ketidaklancaran, tampak bahwa kesalahan pengenalan lazim terjadi pada dua situasi. Pertama, ketika sebelum fragmen atau penggalan unsur wicara, nilai WER cukup tinggi, yakni 28,3%. Juga ketika terjadi pengulangan, kesalahan pengenalan kemungkinan akan terjadi pada salah satu segmen yang diulang. Nilai WER pada bagian repetisi juga cukup tinggi, yakni sekitar 32,3%.

Dari apa yang terlihat pada Tabel 1, jika dibandingkan fitur lainnya, tampaknya ketidaklancaran hanya berpengaruh sedikit pada kesalahan ASR BI. Namun, bisa saja ini terjadi karena lebih dari 60% data dalam bentuk keluaran satu kata. Dengan demikian, agak sulit diukur ketidaklancarannya karena tidak adanya unsur-unsur lain yang muncul menyertai.

Sementara itu, dari fitur kata tunggal atau rangkaian, tampak bahwa kesalahan pada keduanya sama-sama cukup besar, menyentuh tingkat WER di atas 40%, baik untuk keluaran yang berbentuk kata tunggal, maupun keluaran yang berbentuk frase, klausa, maupun kalimat.

Tabel 1 Tingkat kesalahan kata (word error rate = WER) pada data

Fitur	**WER**	**F measure**	**% data**
Sebelum Jeda	16.1	0.62	1.8
Setelah Jeda	15.8	0.643	2.3
Sebelum Fragmen	28.3	0.545	1.8
Setelah Fragmen	21.4	0.582	1.6
Sebelum Repetisi	13.4	0.525	4.7
Setelah Repetisi	14.6	0.567	6.6
Dalam Repetisi	32.3	0.436	16.3
Tunggal	40.0	0.31	64.3
Rangkaian	41.0	0.3	35.7
Kelas Terbuka	43.2	0.3	50.2
Kelas Tertutup	6.7	0.632	41.8
Pemarkah Wacana	31.5	0.37	8.0

Pada fitur kategori kata, berdasarkan hasil analisis, tampaknya ASR BI memang cukup kesulitan untuk mengenali kelas kata terbuka, seperti nomina, verba, dan adjektiva (43.2%). Hal yang sama juga terjadi pada kata-kata pemarkah wacana (31.5%).Namun, tampaknya kelas kata tertutup cukup mudah utuk dikenali oleh ASR. Barangkali itu terjadi karena keanggotaan kelas kata ini cenderung tetap dan tidak pernah berubah. Dengan demikian, model statistik yang digunakan pada ASR mampu memprediksi dan mengantisipasi masukan yang berupa kelas kata tertutup.

5. Kesimpulan

Dari kajian awal tentang ASR BI ini tampak bahwa memang performa sistem aplikasi ini masih cukup jauh dari harapan. Total kesalahannya masih di atas 80%. Kemudian, dari sudut fitur, tampak bahwa faktor ketidaklancaran dalam produksi masukan cukup berpengaruh terhadap kinerja ASR. Sementara itu, faktor yang paling berpengaruh terhadap kinerja ASR BI adalah kategori kelas kata terbuka dan pemarkah wacana. Selain itu, bentuk kata tunggal dan unsur yang lebih panjang daripada kata mempunyai pengaruh yang sama dalam hal kinerja ASR BI.

ASR BI yang dibangun dengan menggunakan model pengetahuan jelas masih jauh dari cukup kemampuannya untuk mengenali dan memroses bunyi ujaran bahasa Indonesia. Pengembangan menggunakan model mesin belajar (*machine learning*) dengan membuat dan menambahkan pangkalan data latihan yang berisi ujaran berbahasa Indonesia tentu akan memperbaiki kinerja ASR BI di masa mendatang. Peningkatan performa dan kinerja ASR BI tentu saja akan

berimbas pada peningkatan penggunaan telepon pintar di Indonesia. ASR BI oleh para pengguna internet dan

REFERENSI

Adda-Decker, M., dan Lamel L. 2005. Do speech recognizers prefer female speakers Dalam Proceeding of INTERSPEECH, 2205-2208.

Dahan, D., Magnuson, J., dan Tanenhaus, M. 2001. Time course of frequency effects in spoken-word recognition: evidence from eye movements. Cognition Psychology 42, 317-367.

Fosler-Lussier, E., dan Morgan, N. 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. Speech Communication 29, 137-158.

Goldwater, S., Jurasfky, D., dan Manning, C.D. 2010. Which word are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Communication 52, 181-200.

Jurafsky, D. dan Martin, J.H. 2008. Speech and Language Processing (2nd Edition). New York: Pearson Prentice Hall.

Luce, P., dan Pisoni, D. 1998. Recognizing spoken words: the neighborhood activation model. Ear Hearing 19, 1-36.

Shinozaki, T., dan Furui, S. 2001. Error analysis using decision trees in spontaneous presentation speech recognition. Dalam Proceeding of ASRU 2001.

Siegler, M., dan Stern, R. 1995. On the effects of speech rate in large vocabulary speech recognition systems. Dalam Proceeding of ICASSP.

Suyanto dan Adityatama, J. 2012. IndoVM: Indonesian Voice Messaging System. Proceedings of 8th International Conference on Information Science and Digital Content Technology (ICIDT), 2012, Volume 1, 145-148.