https://doi.org/10.17576/apjitm-2013-0201-01

http://www.ftsm.ukm.my/apjitm
Asia-Pacific Journal of Information Technology and Multimedia
Jurnal Teknologi Maklumat dan Multimedia Asia-Pasifik
Vol. 2 No. 1, June 2013 : 1-12
e-ISSN: 2289-2192

KORPUS PERTUTURAN SINTAKSIS-PROSODI BAHASA MELAYU

SABRINA TIUN
ROSNI ABDULLAH
TANG ENYA KONG
SITI KHAOTIJAH MUHAMMAD

ABSTRAK

Kertas ini memperihal tentang pembinaan korpus pertuturan Bahasa Melayu untuk diguna dalam pembinaan sistem pertuturan Bahasa Melayu. Korpus pertuturan Bahasa Melayu ini diwakili dengan perwakilan struktur pokok sintaks-prosodi, yang diubah suai daripada struktur perwakilan Structured-String Correspondence (SSTC). Bagi membina korpus pertuturan Bahasa Melayu dalam perwakilan sintaks-prosodi, ayat teks yang sedia kala dalam perwakilan SSTC diguna sebagai skrip rakaman. Melalui rakaman suara berdasarkan skrip tersebut, fitur prosodi diekstrak keluar dan dianotasi pada struktur pokok SSTC, dan pada masa yang sama, fail bunyi dipaut pada nod struktur pohon SSTC. Pada akhir pemprosesan rakaman dan anotasi, mini korpus pertuturan yang diwakili dengan perwakilan sintaksis-prosodi yang mengandungi 422 ayat, 1720 frasa dan 6978 unit perkataan berjaya dihasil.

Kata kunci: Korpus pertuturan, Sistem pertuturan Bahasa Melayu, Sintaks-Prosodi, Struktur perwakilan sintaks-prosodi, sintaks, prosodi.

ABSTRACT

This paper presents a construction of a Malay syntax-prosody speech corpus that was intentionally built for a Malay concatenative speech synthesizer system. The Malay syntax-prosody speech corpus was represented by the syntax-prosody tree structures representation, an adaptation of a syntactic tree structure representation called Structured-String Tree Correspondence (SSTC). In order to build the Malay speech corpus in syntax-prosody representation, existing sentences of SSTC were used for recording. Prosodic features, which were obtained from the recording files, were annotated to the SSTC tree structure and at the same time, sound files were linked to the SSTC trees. The speech corpus contains 422 sentences, 1720 phrase and 6978 word units.

Keywords: Speech Corpus, Malay Speech Synthesis System, Syntax-Prosody, Syntax-Prosody Structure Representation, Syntax, Prosody

PENGENALAN

Sistem sintesis pertuturan yang ideal sepatutnya anjal tanpa ada pengurangan kualiti dari segi kecerdasan dan bunyi semulajadinya. Maksud istilah keanjalan ini ialah keupayaan sesebuah sistem sintesis pertuturan untuk mensintesis sebarang input teks. Namun, Yi dan Glass (1988), dan juga Klabbers (2000) menyatakan keanjalan serta tabii adalah sesuatu nilai kualiti yang saling berlawanan dalam membina sebuah sistem sintesis pertuturan. Sekiranya sebuah sistem sintesis pertuturan dibina dengan mengutama kualiti keanjalan, maka kualiti semulajadinya mungkin tidak dapat dibina pada tahap yang tinggi, dan begitu juga sebaliknya. Pada keadaan semasa, kebanyakan sistem sintesis pertuturan dibina dengan mengutama keanjalan (Klabbers 2000; Cox et al 2000). Namun, sistem sintesis pertuturan boleh juga dibina dengan mengutama kualiti semulajadinya terlebih dahulu.

Tujuan penyelidikan ini ialah membina korpus pertuturan bahasa Melayu yang diwakili dengan perwakilan sintesis-prosodi. Korpus ini diguna untuk membina model sintesis

pertuturan yang berbunyi semulajadi berserta dengan fitur keanjalan yang minimum. Dalam kajian sintesis pertuturan, jenis pembinaan korpus dan sistem sintesis pertuturan berkait-rapat, terutamanya pada kaedah sintesis pertuturan yang berdasarkan korpus.

Pendekatan yang diguna bagi membina sistem sintesis pertuturan ialah kaedah berdasarkan korpus iaitu mengadaptasi kerangka kerja mesin terjemahan UTMK yang berdasarkan contoh (Ye 2006) dan korpus pertuturan diwakili dengan perwakilan struktur sintaks-prosodi. Dalam korpus pertuturan sintaks-prosodi, setiap ayat dalam korpus diwakili oleh satu pokok kepautan sintaks-prosodi. Setiap nod pada pokok tersebut dianotasi dengan fitur prosodi dan juga diselaras dengan seunit bunyi perkataan. Oleh itu, apabila ayat baharu disintesis mengguna sistem pertuturan ini, ayat input tersebut terlebih dahulu dihurai pada struktur pokok sintaks-prosodi dan kemudiann semua unit bunyi yang terdapat pada nod pokok digabung untuk menghasil bunyi pertuturan pada ayat yang dimasukkan.

Kami memilih untuk mewakili korpus pertuturan dalam struktur pokok sintaks-prosodi berdasarkan penyataan Mobious (2000) iaitu unit pertuturan sesuai dipilih berdasarkan konteks. Dalam kajian ini, pemilihan berdasarkan konteks bermaksud memilih unit pertuturan dengan mengadaptasi sistem penghurai berdasarkan-contoh dalam mesin terjemahan UTMK (Ye 2006). Pemilihan unit pertuturan dilakukan secara tersirat dengan membina pokok perwakilan sintaks-prosodi pada input ayat yang hendak disintesis. Oleh kerana nod pokok diselaras dengan unit bunyi pertuturan, maka unit bunyi pada pokok yang dibina secara langsung dianggap merupakan unit pertuturan yang sesuai untuk ayat sasaran. Dalam erti kata lain, jika unit pertuturan tersebut dicantum, maka tidak dapat dibeza sama ada bunyi pertuturan tersebut ditutur oleh manusia ataupun mesin.

Penggunaan penghurai berdasarkan contoh dan perwakilan sintaks-prosodi dalam sistem pertuturan adalah bagi mengelak ketidakpadanan prosodi pada cantuman unit pertuturan yang terpilih. Manakala, untuk mengelak ketidakpadanan penggalan pada cantuman unit pertuturan, saiz unit pertuturan yang dipilih ialah saiz yang mempunyai kurang impak artikulasi dengan fonem yang berjiranan. Oleh yang demikian, dalam korpus pertuturan, hanya tiga jenis unit pertuturan yang dipilih, iaitu: frasa, perkataan dan sub perkataan (Tiun et al. 2011).

SINTAKS-PROSODI DAN PERWAKILAN STRUKTUR POKOK DALAM SINTESIS PERTUTURAN

Penghuraian sintaks dan prosodi menyahtaksa antara satu dengan yang lain. Contohnya, penghurai prosodi selalunya diguna untuk menyahtaksa struktur sintaks ayat dalam sistem pengecaman pertuturan atau sistem pemahaman pertuturan (Hunt 1992; Bear & Price 1990; Gallwitz et al. 2002). Manakala, penghurai sintaks pula selalunya diguna dalam menyahtaksa permasalahan prosodik dalam sistem sintesis pertuturan (Bachenko & Fitzpatrick 1900; Atterer & Klein 2002).

Penggunaan penghurai sintaksis untuk menyahtaksa prosodi dilakukan secara meluas dalam sistem sintesis pertuturan yang mengguna pendekatan cantuman tetap. Ini kerana fitur prosodi diguna untuk mengubah sifat akustik pada unit pertuturan bagi sasaran ayat yang hendak disintesis. Struktur frasa prosodi ialah permasalahan ketaksaan yang selalu dinyahtaksa dengan penghurai sintaksis; sebagai contoh, dalam kerja penyelidikan Bachenko dan Fitzpatrick (1990), maklumat perkataan seperti leksikal dan golongan kata diguna untuk menentu hentian frasa, iaitu salah satu fitur prosodi. Pendekatan Bachenko dan Fitzpatrick (1990) hanya mengguna maklumat perkataan pada ayat dan tidak pada struktur sintaksis ayat tersebut. Blin dan Miclet (2000) berpendapat jika maklumat tentang struktur sintaksis ayat diambil kira, penyahtaksaan fitur prosodi menjadi tepat kerana banyak maklumat diguna dalam proses penyahtaksaan.

Blin dan Miclet (2000) dan Taylor (2000) mengguna keseluruhan struktur pokok ayat sasaran untuk menyahtaksa prosodi bagi mendapatkan unit-unit sintesis pertuturan yang

mempunyai fitur prosodi yang paling sesuai. Dalam kajian Blin dan Miclet (2000), bagi menyahtaksa prosodi ayat sasaran, ayat tersebut dihurai dalam struktur pokok *performance.* Pokok *performance* tersebut seterusnya dibanding dengan pokok *performance* dalam korpus pertuturan. Oleh yang demikian, fitur prosodi pokok yang sepadan dari korpus pertuturan dianggap fitur prosodi yang sesuai untuk ayat sasaran. Pokok *performance* Blin dan Miclet (2000) ialah kombinasi antara struktur pokok *performance* (bahagian atas) dan struktur pokok sintaksis (bahagian bawah). Pokok *performance* merupakan pokok yang cabangnya terbentuk berdasarkan kekuatan hentian antara perkataan. Manakala, pokok sintaksis pada bahagian bawah, dilanjutkan dalam struktur fonologi. Lanjutan struktur fonologi tersebut ialah nod terminal dalam bentuk fonem, yang juga merupakan anak nod pada nod suku kata.

Dalam kajian Taylor (2000), pertuturan sintetik dihasil dengan memadan struktur pokok sasaran dengan struktur pokok dalam pangkalan data pertuturan. Struktur pokok ialah gabungan pokok *metrical* (pada bahagian atas) dan struktur pokok fonologi sub-suku kata (pada bahagian bawah). Asalnya, pokok gabungan tersebut ialah pokok sintaksis, dengan mengguna rumus linguistik, bahagian atas pokok (iaitu di atas nod perkataan) struktur pokok diubah menjadi struktur pokok *metrical,* dengan nod ditandai dengan label ‘s’ untuk hentian antara dua perkataan yang jelas dan label ‘w’ untuk hentian antara dua perkataan yang tidak begitu jelas. Manakala di bawah nod perkataan, struktur pokok fonologi dijana dengan menjadikan nod fonem sebagai nod terminal.

Kebanyakan pakar linguistik bukan sahaja berpendapat struktur sintaks tidak sinonim atau isomofik dengan struktur prosodi, malah tidak jelas hubungan padanan di antara kedua-dua struktur tersebut. Namun begitu, kekaburan hubungan struktur prosodi dan sintaks boleh dimanipulasi untuk menyelesai permasalahan ketaksaan di antara satu dengan yang lain. Dalam kajian ini, penghurai sintaks diguna untuk mendapat unit pertuturan yang sesuai untuk disintesis menjadi bunyi pertuturan.

GAMBARAN KESELURUHAN HUBUNGAN SINTAKS-PROSODI BAHASA MELAYU

Rujukan mengenai prosodi bahasa Melayu secara amnya sukar didapati. Sesetengah isu seperti sama ada bahasa Melayu adalah bahasa *syllable-timed*, atau sama ada mempunyai tekanan atau tidak, masih terus dibincang. Don et al. (2008) menyatakan bahasa Melayu bukan bahasa *syllable-timed* kerana jarak masa suku kata bahasa Melayu adalah tidak sekata selain daripada Bahasa Melayu merupakan bahasa yang tidak mempunyai tekanan. Asas penyataan tersebut adalah berdasarkan pemerhatian corak parameter tempoh dan F0 pada 111 jumlah perkataan kata isi. Namun, kajian Kassin (2000) menunjukkan sebaliknya apabila secara tidak sengaja menunjukkan bahasa Melayu merupakan bahasa yang mempunyai tekanan iaitu tekanan primer dan tekanan sekunder. Tekanan primer ialah tekanan yang wujud pada suku kata yang kedua terakhir, manakala tekanan sekunder adalah tekanan pada suku kata pertama. Contoh, dalam perkataan bidadari (*/bi.da.da.ri/*), suku kata /*da*/ menerima tekanan primer dan suku kata pertama iaitu /*bi*/ menerima tekanan sekunder (Kassin, 2000). Namun, corak tekanan perkataan berubah jika perkataan bukan kata akar atau perkataan mengandungi vokal tengah.

Kajian Payne (1970) menerangkan intonasi bahasa Melayu berkait rapat dengan jenis ayat dan kedudukan frasa, sama ada pada kedudukan permulaan, tengah atau penghujung. Pembahagian ayat pada frasa ditandai dengan hentian yang dinamai hentian tergantung (*suspense break*). Menurut Payne (1970), hentian tergantung atau hentian frasa boleh dikenal pasti jika ada fitur jeda (*pause*) atau dalam bunyi vokal atau konsonan yang panjang. Bahasa Melayu, intonasi boleh menandai konstituen sintaksis, dengan intonasi menaik dan selepas itu menurun menunjukkan sempadan subjek dan predikat. Penyataan sedemikian turut diberi oleh Abdul-Wahab (1988) yang menyatakan intonasi berhubung kait dengan struktur sintaksis. Justeru, dengan hanya mendengar intonasi percakapan seseorang, maka penilaian boleh

dilakukan sama ada penutur tersebut mengetahui sintaks bahasa Melayu atau sebaliknya. Dalam erti kata lain, dalam bahasa Melayu struktur sintaks dan struktur prosodi adalah berkait rapat.

Kajian Abdul-Wahab (1988) mengkategori nada intonasi bahasa Melayu pada empat aras; dengan '1' sebagai aras yang paling rendah dan '4' sebagai aras yang teratas. Jadual 1 menunjukkan corak intonasi berdasarkan jenis ayat: (a) untuk ayat deklaratif, corak intonasi selalunya dalam corak 2-3-2-3, dan (b) ayat tanya yang mempunyai corak intonasi 2-4-3-4, 2-4-2-4 atau 2-4-1, manakala (c) ayat seru mempunyai corak sama ada 2-3 atau 2-3-1.

Corak intonasi yang dinyatakan oleh Abdul-Wahab (1988) dikaji selanjutnya oleh Zahid dan Shah Omar (2006) dengan pendekatan eksperimental, iaitu pemerhatian corak intonasi mengguna alat bantu penglihatan. Dalam kajian experimental tersebut, aras 1 diganti dengan label 'L' dan aras 4 sebagai 'H', dan daripada hasil pemerhatian, Zahid dan Shah Omar (2006) mendapati semua corak intonasi boleh diterima kecuali intonasi ayat deklaratif 2-3-2-3 yang janggal dan tidak dapat dipadankan pada struktur sintaks bahasa Melayu.

JADUAL 1. Senarai jenis ayat dan corak intonasi (Abdul-Wahab, 1988)

Jenis Ayat	Corak intonasi
Deklaratif	2-4-2-3
Tanya	2-4-3-4
	2-4-2-4
	2-4-1
Seru	2-3
	2-3-1

Berdasarkan perbincangan mengenai prosodi bahasa Melayu ini, dapat difahami mengapa bahasa Melayu yang dianggap sebagai bahasa yang mudah oleh pakar linguistik tetapi kerja penyelidikan ke atas bahasa Melayu agak sukar. Kekeliruan sama ada bahasa Melayu mempunyai tekanan atau tidak, atau pendapat yang berbeza-beza mengenai corak intonasi bahasa Melayu memberi alasan yang kuat untuk sintesis pertuturan bahasa Melayu dibina berdasarkan korpus.

PEMBINAAN KORPUS PERTUTURAN SINTAKS-PROSODI

Dalam pembinaan korpus pertuturan sintaks-prosodi, fitur prosodi yang diesktrak dari korpus pertuturan dianotasi pada struktur pokok sintaks. Gabungan dua pengetahuan ini, pokok sintaksis dan anotasi prosodi, dinamai perwakilan sintaks-prosodi. Dalam perwakilan sintaks-prosodi tersebut, fitur prosodi iaitu frasa (*phrasal*) dan kelantangan bunyi (*prominence*) dianotasi pada struktur pergantungan pokok sintaksis. Kajian ini mengguna keseluruhan struktur pokok perwakilan untuk memilih calon sasaran unit pertuturan yang sesuai untuk digabung menjadi ayat sintesis pertuturan. Penerangan terperinci mengenai pemilihan calon unit pertuturan boleh dirujuk dengan mendalam dalam Tiun et al (2012).

Unit pertuturan korpus pertuturan bahasa Melayu dalam kajian ini ialah ayat, frasa, perkataan dan sub-perkataan. Setiap ayat dalam korpus diwakili dengan struktur pokok sintaks-prosodi dan diindeks berdasarkan struktur anotasi *Structured String Tree Correspondence (SSTC)*. SSTC bertindak sebagai alat yang menganotasi kesepadanan di antara teks ayat dengan perwakilan struktur pokok sintaks-prosodi. Seksyen 2.1 seterusnya memberi penerangan ringkas tentang SSTC dan dalam seksyen 2.2, diterangkan proses pembinaan korpus pertuturan sintaks-prosodi.

STRUCTURED STRING TREE CORRESPONDENCE (SSTC)

SSTC atau *Structured String Tree Correspondence* ialah struktur anotasi yang mengandungi untaian (*string/language*), struktur pokok untaian (*tree*) dan kesepadanan (*corresponding*) di

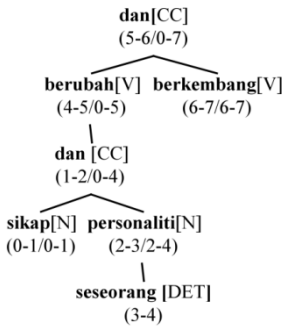
antara kedua-duanya. Skema anotasi SSTC mendefinisi kesepadanan di antara untaian dengan struktur pokoknya berdasarkan dua kesepadanan yang saling berkait rapat, iaitu: (1) kesepadanan antara nod dan sub-untaian dan (2) kesepadanan di antara sub-pokok dan sub-untaian. Kedua-dua kesepadanan yang saling berkait rapat itu dikod bersama dengan selangan (*intervals*) yang dinamai SNODE dan STREE. SNODE ialah selangan bagi kesepadanan sub-untaian pada pokok nod. Manakala STREE ialah selangan kesepadanan di antara sub-untaian dengan sub-pokok. Dalam Rajah 1, SNODE-SNODE pokok tersebut adalah seperti dalam Jadual 2. Manakala STREE-STREE disenarai dalam Jadual 3.

JADUAL 2. Senarai SNODE-SNODE bagi pokok dalam Rajah 1

Bil	Nod	Selangan
1	*sikap[N]*	(0-1)
2	*dan[CC]*	(1-2)
3	*personaliti[N]*	(2-3)
4	*seseorang[DET]*	(3-4)
5	*berubah[V]*	(4-5)
6	*dan[CC*	(5-6)
7	*berkembang[V]*	(6-7)

JADUAL 3. Senarai STREE-STREE pokok bagi Rajah 1

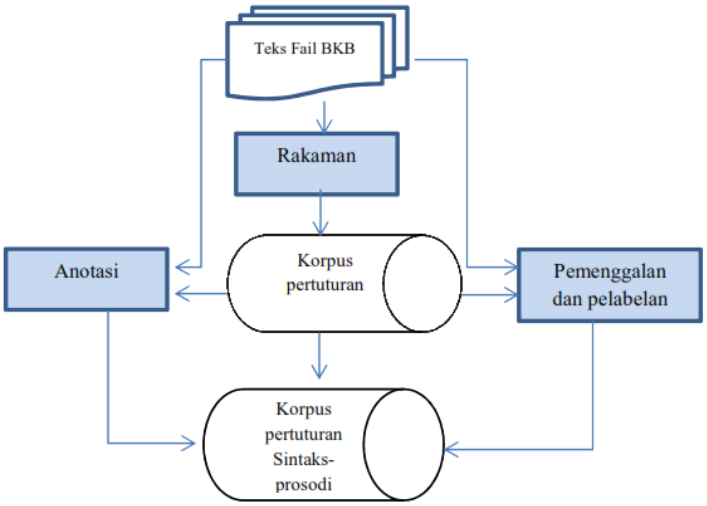
Bil	Untaian/Sub-untaian	Nod akar sub-pokok yang sepadan	Selangan
1	*'sikap dan personaliti seseorang berubah dan berkembang'*	*dan[CC]*	(0-8)
2	'*sikap dan personaliti seseorang berubah'*	*berubah[V]*	(0-5)
3	'*sikap dan personaliti seseorang'*	*dan[CC]*	(0-4)
4	*'personaliti seseorang'*	*personaliti[N]*	(2-4)
5	*'berkembang'*		(6-7)
6	*'sikap'*		(0-1)



RAJAH 1. Pokok kepautan yang dikod dengan anotasi SSTC untuk untaian *sikap$_1$ dan $_2$ personaliti$_3$ seseorang$_4$ berubah$_5$ dan$_6$ berkembang$_7$*

PENYEDIAAN KORPUS PERTUTURAN SINTAKS-PROSODI

Ringkasan rajah pada Rajah 2 menggambar proses pembinaan korpus pertuturan bahasa Melayu (kotak yang berwarna gelap), iaitu: (1) Proses rakaman, (2) proses anotasi dan (3) proses pemenggalan dan pelabelan.

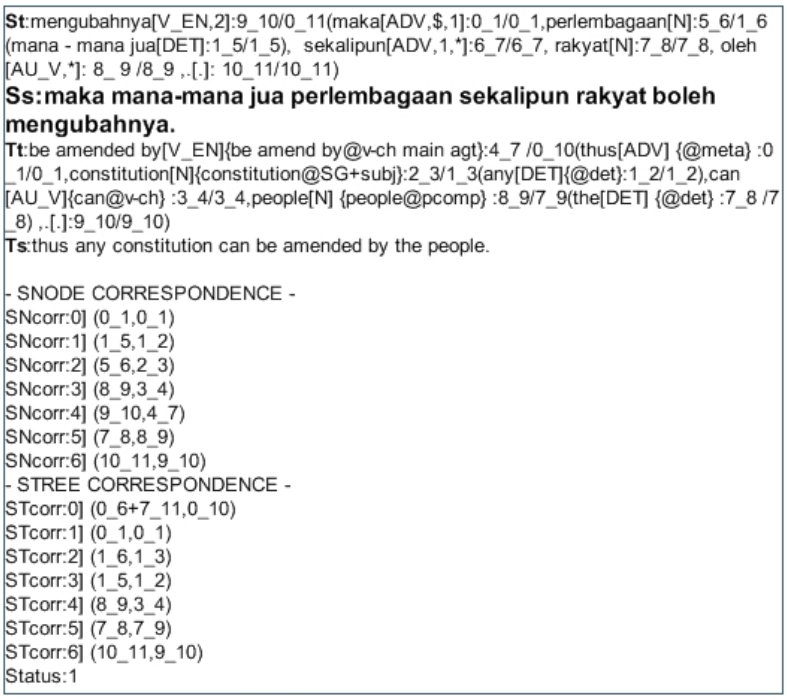


RAJAH 2. Gambaran keseluruhan proses penyediaan korpus pertuturan sintaks-prosodi.

PROSES RAKAMAN

Skrip rakaman korpus pertuturan ialah koleksi ayat teks yang diekstrak keluar secara automatik daripada pangkalan pengetahuan dwibahasa mesin terjemahan bahasa Melayu-Inggeris, atau dikenali sebagai BKB. Fail dari pangkalan pengetahuan tersebut mengandungi maklumat kejajaran di antara ayat bahasa Melayu dan ayat bahasa Inggeris dalam anotasi struktur SSTC yang segerak. Bagi tujuan rakaman, hanya ayat bahasa Melayu diambil kira dan diekstrak keluar. Rajah 3 menunjukkan ayat *maka mana-mana jua perlembagaan sekalipun rakyat boleh mengubahnya* yang ditebalkan, dimasukkan ke dalam skrip rakaman. Terdapat 422 ayat bahasa Melayu yang diekstrak keluar dari fail BKB domain JAKIM. Fail domain JAKIM bermaksud fail teks yang mempunyai topik berkaitan dengan agama Islam.

Menurut Clark et al. (2004), ciri asas yang sepatutnya ada pada pembaca skrip rakaman ialah; pembaca mempunyai artikulasi yang jelas serta mampu membaca skrip yang panjang secara konsisten dalam gaya yang semulajadi. Dalam kajian ini, pembaca rakaman adalah seorang yang mempunyai artikulasi yang jelas, mempunyai pengalaman merakam suara untuk korpus pertuturan sebelum ini dan beliau berbangsa Melayu. Semua kerja rakaman dilakukan di Makmal Akustik UTMK, iaitu makmal yang merupakan bilik kedap bunyi yang sudah banyak kali diguna sebagai bilik rakaman untuk menghasil korpus pertuturan. Jadual 4 menunjukkan spesifikasi peralatan yang diguna semasa kerja rakaman dijalankan.



RAJAH 3. Format fail BKB

Proses rakaman dilakukan dalam empat sesi, yang mana secara puratanya 100 ayat dapat dirakam pada setiap sesi. Sebelum setiap sesi rakaman bermula, kecuali sesi rakaman yang pertama, suara pembaca perlu dirakam untuk tujuan penanda aras yang diguna untuk memasti kekuatan bunyi suara dapat diselaras bagi semua sesi rakaman. Proses penyelarasan ini dilakukan dengan mendengar dan membanding sesi bunyi rakaman yang terdahulu dengan yang terkini. Persediaan peralatan rakaman dipasti sama pada semua sesi rakaman. Jika semasa proses rakaman, pembaca membuat kesilapan pembacaan, proses rakaman tidak dihenti, tetapi pembaca perlu membaca bahagian yang salah sekali pada penghujung proses rakaman pada satu-satu masa.

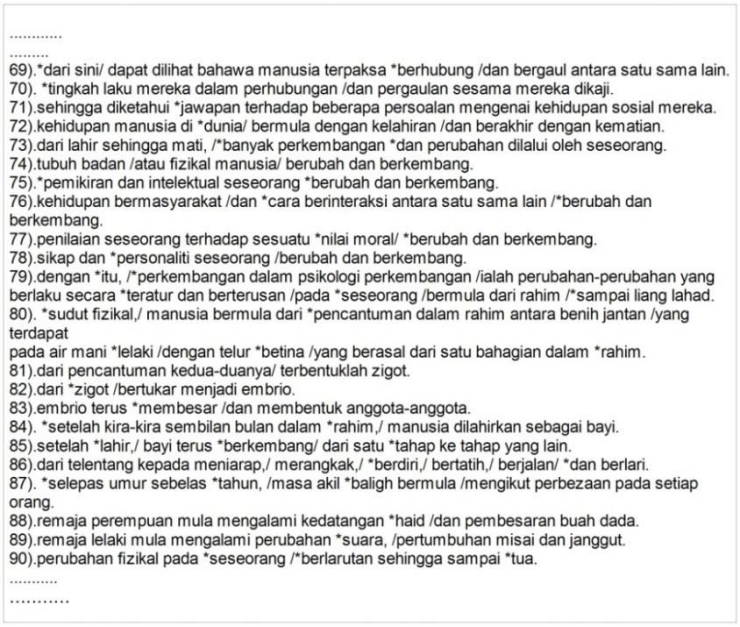
Mikrofon disambung secara langsung pada komputer riba, oleh itu kebanyakan sumber bunyi bising; litar elektrik dan bunyi kipas dari komputer riba, turut serta dirakam semasa sesi rakaman. Bunyi bising dibersih mengguna alat bantu *noise reduction* dengan perisian *Adobe Audition*. Selepas itu rakaman yang dibersih dari bunyi bising disimpan dalam format 44 *khz, mono* dan bentuk fail *.wav.* Hasil dari proses rakaman sebanyak empat sesi rakaman, korpus pertuturan yang dihasil mempunyai saiz sebesar tujuh jam dengan purata satu minit per ayat. Speksifikasi proses yang terperinci boleh dicapai pada pautan http://www.ftsm.ukm.my/MalaySpeechCorpus.

JADUAL 4. Spesifikasi peralatan untuk proses rakaman

Peralatan	Spesifikasi
Mikrofon *Pre-amp*	*Samson C01U-based*
CPU	*Toshiba laptop (model 'protégé')*
Headset	*AK model K.66*
Perisian rakaman	*Adobe Audition 2.0*

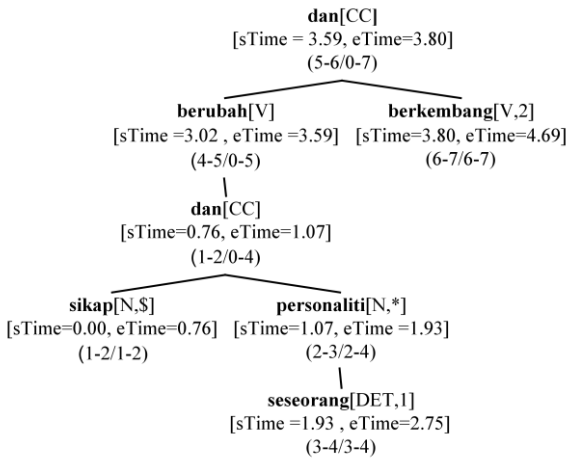
PROSES ANOTASI

Korpus pertuturan dalam kajian ini dianotasi dengan fitur prosodik yang terdiri dari empat jenis; kelantangan bunyi, jenis hentian frasa dan kedudukan perkataan dalam ayat. Secara manual, penganotasi fitur prosodik, iaitu seorang penutur bahasa Melayu yang asli, mendengar rakaman dan menanda perkataan yang mempunyai kelantangan bunyi dengan simbol '*'. Bagi jenis hentian, cuma dua jenis hentian diambil kira: Hentian frasa dan hentian sempadan yang lengkap. Secara manual juga, penganotasi menanda hentian frasa yang jelas menggunakan simbol '/'. Bagi hentian sempadan lengkap, penganotasi tidak perlu berbuat apa-apa kerana simbol titik '.' sudah cukup diguna sebagai panduan. Rajah 4 menunjukkan contoh teks rakaman yang telah dianotasi dengan fitur prosodi. Teks rakaman yang dianotasi diguna sebagai panduan untuk menganotasi pokok SSTC dengan maklumat prosodik. Pada pokok SSTC, nod yang mempunyai kelantangan bunyi dianotasi dengan simbol '*'. Nod yang mempunyai informasi hentian frasa dianotasi dengan simbol '1', dan hentian pada sempadan yang lengkap dianotasi dengan simbol '2'. Nod yang mengandungi leksikal pada kedudukan pertama dalam ayat ditanda dengan simbol '$'.



.....
.....
69).*dari sini/ dapat dilihat bahawa manusia terpaksa *berhubung /dan bergaul antara satu sama lain.
70). *tingkah laku mereka dalam perhubungan /dan pergaulan sesama mereka dikaji.
71).sehingga diketahui *jawapan terhadap beberapa persoalan mengenai kehidupan sosial mereka.
72).kehidupan manusia di *dunia/ bermula dengan kelahiran /dan berakhir dengan kematian.
73).dari lahir sehingga mati, /*banyak perkembangan *dan perubahan dilalui oleh seseorang.
74).tubuh badan /atau fizikal manusia/ berubah dan berkembang.
75).*pemikiran dan intelektual seseorang *berubah dan berkembang.
76).kehidupan bermasyarakat /dan *cara berinteraksi antara satu sama lain /*berubah dan berkembang.
77).penilaian seseorang terhadap sesuatu *nilai moral/ *berubah dan berkembang.
78).sikap dan *personaliti seseorang /berubah dan berkembang.
79).dengan *itu, /*perkembangan dalam psikologi perkembangan /ialah perubahan-perubahan yang berlaku secara *teratur dan berterusan /pada *seseorang /bermula dari rahim /*sampai liang lahad.
80). *sudut fizikal,/ manusia bermula dari *pencantuman dalam rahim antara benih jantan /yang terdapat
pada air mani *lelaki /dengan telur *betina /yang berasal dari satu bahagian dalam *rahim.
81).dari pencantuman kedua-duanya/ terbentuklah zigot.
82).dari *zigot /bertukar menjadi embrio.
83).embrio terus *membesar /dan membentuk anggota-anggota.
84). *setelah kira-kira sembilan bulan dalam *rahim,/ manusia dilahirkan sebagai bayi.
85).setelah *lahir,/ bayi terus *berkembang/ dari satu *tahap ke tahap yang lain.
86).dari telentang kepada meniarap,/ merangkak,/ *berdiri,/ bertatih,/ berjalan/ *dan berlari.
87). *selepas umur sebelas *tahun, /masa akil *baligh bermula /mengikut perbezaan pada setiap orang.
88).remaja perempuan mula mengalami kedatangan *haid /dan pembesaran buah dada.
89).remaja lelaki mula mengalami perubahan *suara, /pertumbuhan misai dan janggut.
90).perubahan fizikal pada *seseorang /*berlarutan sehingga sampai *tua.
.....
.....

RAJAH 4. Korpus pertuturan yang dianotasi

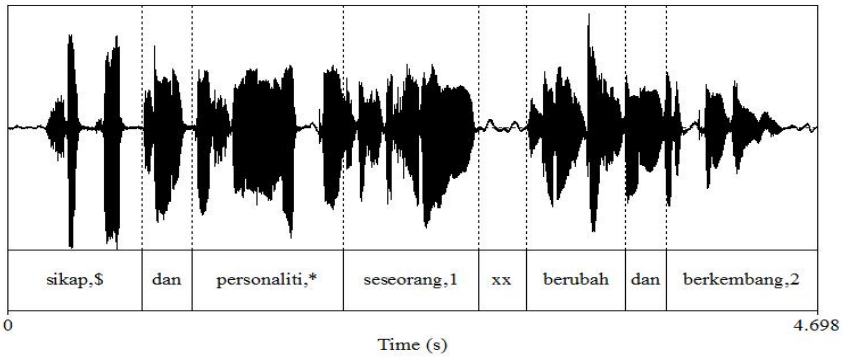


RAJAH 5. Pokok sintaks-prosodi yang dikod dengan skema SSTC.

Semua perkataan yang ditandai dengan simbol prosodik dianggap mempunyai tahap ciri-ciri akustik yang jauh berbeza daripada perkataan yang lain dalam ayat yang sama. Oleh kerana itu, perkataan tersebut harus dibeza dengan perkataan yang lain. Rajah 5 merupakan contoh bagaimana nod dalam pokok sintaks-prosodi dikod dalam skema SSTC dan fitur prosodik diselaras dengan penggalan pertuturan dari fail bunyi korpus pertuturan dengan fail bunyi yang mempunyai untaian '*sikap,$ dan personaliti,* seseorang,1 berubah dan berkembang,2'.*

PROSES PEMENGGALAN DAN PELABELAN

Pemenggalan unit pertuturan dalam korpus pertuturan dilakukan secara semi-automatik. Memandangkan satu fail adalah padanan pada satu ayat, oleh itu korpus pertuturan mengandungi sebanyak 422 fail *.wav.* Setiap fail *.wav* tersebut dipotong ke unit frasa dan perkataan. Kedua-dua unit frasa dan perkataan dipotong secara automatik dengan perisian *Praat*, dan selepas itu potongan unit tersebut diperiksa dan disunting secara manual. Selepas itu, potongan pertuturan dilabel dengan huruf ortografik dan simbol prosodik. Semua skrip *Praat* yang diguna dalam kajian ini boleh diperoleh dari laman sesawang http://www.helsinki.fi/lennes/praat-scripts.



RAJAH 6. Penggalan pertuturan dari fail *.wav* bagi untaian yang beranotasi *sikap,$ dan personaliti,* seseorang,1 berubah dan berkembang,2*

Memandangkan kedua-dua proses pelabelan dan pemenggalan unit pertuturan mengguna perisian *Praat*, oleh itu semua maklumat disimpan dalam fail berformat *.textgrid*. Rajah 6 ialah contoh fail yang dipenggal dan dilabel bagi unit pertuturan perkataan.

HASIL

Selepas ketiga-tiga proses dijalankan; (1) rakaman, (2) anotasi, dan (3) pemenggalan dan pelabelan, korpus pertuturan sintaks-prosodi tersebut diindeks secara automatik untuk memudah proses dapatan unit pertuturan semasa proses sintesis nanti. Korpus pertuturan sintaks-prosodi ini diindeks berdasarkan struktur pokok sintaks-prosodi, iaitu: keseluruhan pokok sepadan dengan unit ayat, sub-pokok sepadan dengan unit frasa dan nod sebagai unit perkataan. Oleh kerana pemenggalan unit frasa bergantung pada simbol '1', maka pangkalan data terindeks sintaks-prosodi hanya menyimpan sub-pokok yang sepadan dengan unit frasa yang dipenggal berdasarkan simbol '1'.

Hasil daripada proses rakaman, anotasi, pemenggalan dan pelabelan dan pengindeksan dalam pangkalan data, terhasil mini korpus pertuturan bahasa Melayu yang mengandungi 422 unit ayat, 1720 unit frasa dan 6787 unit, seperti dalam Jadual 5. Capaian untuk skrip rakaman

yang dianotasi dan spesifiksi rakaman boleh dicapai pada pautan http://www.ftsm.ukm.my/MalaySpeechCorpus .

JADUAL 5. Ringkasan maklumat mengenai unit pertuturan yang terkandung dalam korpus pertuturan sintaks-prosodi Bahasa Melayu

Jenis unit pertuturan	Jumlah unit
Ayat	422
frasa	1720
perkataan	6787

PERBINCANGAN

Sistem pertuturan yang mengguna korpus pertuturan sintaks-prosodi bahasa Melayu ini dinamai sistem UTMK-MSS, iaitu *Unit Terjemahan Melalui Komputer-Malay Speech Synthesis.* Penerangan lanjut mengenai bagaiman korpus ini diguna-untuk proses sintesis boleh diperoleh dari Tiun et al. (2012). Dalam Tiun et al. (2012), suara sintesis sistem UTMK-MSS dibanding dengan suara sintesis dua jenis sistem sintesis bahasa Melayu yang berlainan dan juga suara semulajadi (rakaman). Berdasaran penilaian subjektif *Mean Opinion Scores* (MOS) yang mengguna penilaian manusia, suara sintektik dari sistem UTMK-MSS ternyata berbunyi semulajadi, dengan jumlah purata keseluruhan skor MOS 3.3, berbanding dengan ketiga-tiga sistem sintesis pertuturan bahasa Melayu yang hanya memperolehi purata keseluruhan skor MOS sebanyak 1.95 dan 1.80. Jika dibanding dengan suara semulajadi manusia (rakaman) dengan MOS skor 4.75 dan skor MOS sistem UTMK-MSS 3.35, ternyata kualiti suara semulajadi UTMK-MSS lebih rendah berbanding suara semulajadi manusia. Namun, sekiranya diambil kira purata piawaian keseluruhan MOS skor untuk sistem sintesis, iaitu dalam julat 2.5 - 3.5, kualiti semulajadi suara sintektik UTM-MSS berada dalam nilai yang baik.

Secara ringkasnya, penyediaan korpus pertuturan memerlukan penelitian dari segi jenis pendekatan sistem sintesis, jenis fitur yang perlu diambil kira dan bagaimana struktur korpus pertuturan tersebut. Manakala dari perincian mengenai spesifikasi alatan, bilik, prosedur and perisian rakaman harus diterangkan dengan jelas bagi memboleh kerja kajian ini dapat diulang oleh pengkaji lain. Bagi kajian dan penyelidikan akan datang, kami melihat korpus pertuturan dapat dikembang dalam tiga jenis perspektif: (1) Korpus ini ditambah saiznya agar sasaran sistem sintesis pertuturan mempunyai rangkuman jumlah unit pertuturan yang banyak supaya masalah ketiadaan unit pertuturan yang sesuai dapat diatasi, (2) membina alat pemperosesan automatic bagi semua kerja yang dijalankan secara manual atau semi-automatik, dan (3) korpus pertuturan ini adalah lebih baik jika dilengkapi dengan informasi fitur akustik prosodi seperti F0, tenaga dan dipenggal pada penggalan unit yang kecil seperti fonem, agar korpus ini boleh juga diguna untuk sistem sintesis pertuturan kaedah cantuman tetap.

PENGHARGAAN

Penulis berterima kasih pada Anuar Mansor dan Norliza Hani atas sumbangan dalam penyediaan korpus pertuturan sintaks-prosodi bahasa Melayu ini.

RUJUKAN

Abdul-Wahab, A. 1988. *Intonasi dalam Hubungan dengan Sintaksis Bahasa Melayu*. Kuala Lumpur:Sasbadi Sdn.Bhd.

Abney, S. 1992. Prosodic structure, performance structure and phrase structure. In Mitchell P. Marcus, M. (ed) *Proceeding of 5th DARPA workshop on Speech and Natural Language*. San Mateo: Morgan Kaufmann Publishers, 425-428.

Al-Adhaileh, M. H. 2002. *Synchronous Structured String-Tree Correspondence (S-SSTC) And Its Application for Machine Translation*. PhD thesis, Universiti Sains Malaysia, Pulau Pinang.

Blin, L. & Miclet, L. 2000. Generating synthetic speech prosody with lazy learning in tree Structure. In Cardie, C., Daelemans, W., Nédellec, C. and E. Tjong Kim Sang (eds.) *Proceedings of CoNLL-2000 and LLL-2000*. New York: Association for Computational Linguistics, 87–90.

Don, Z. M., Knowles, G., & Janet, Y. 2008. How Words Can Be Misleading: A Study of Syllable Timing and "stress" in Malay. *The Linguistics Journal*, 3(2): 66-8.

Heldner, M. & Megyesi, B. 2003. Exploring the prosody-syntax interface in conversions. In D. Recasens, M. J. Solé and J. Romero (eds) *Proceedings of 15th ICPhS, XV Intl Conference of Phonetic Sciences*. Barcelona: 15th *ICPhS*, 2501-2504.

Hirschberg, J. & Rambow, O. 2001. Learning prosodic features using a tree representation. In P. Dalsgaard, B. Lindberg, H. Benner and Z. H. Tan (eds.) *Proceedings of the Eurospeech 2001*. Baixas, France: ISCA, 1175-1178.

Kassin, T. A. (2000). *The Phonological Word in Standard Malay*. PhD thesis, University of Newcastle, Newcastle Upon Tyne.

Klabbers, E. & Veldhuis, R.. 2001. Reducing audible spectral discontinuities. *IEEE Transcations on Speech and Audio Processing*, 9(1):39–51.

Maris, Y. 1980. *The Malay Sound System.* Kuala Lumpur: Fajar Bakti.

Möbius, B. 2000. Corpus-based speech synthesis: Methods and challenges. *Forum phoneticum*, 69:79–96.

Payne, E. 1970. *Basic Syntactic Structures in Standard Malay*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Schafer, A. J. 1997. *Prosodic parsing: the role of prosody in sentence comprehension*. PhD thesis, University of Massachusetts, Amherts.

Selkirk, E. 2008. Chapter 25: The Prosodic Function Word in *Optimality Theory in Phonology*. United Kingdom: Blackwell Publishing Ltd.

Stöber, K., Portele, T., Wagner, P., and Hess, W. 1999. Synthesis by word concatenation. In V. Obran, G. Olaszy, G. Nemeth and K. Erdohegy (eds.) *EUROSPEECH'99*. Bonn: European Speech Communication Association, 619-622.

Steedman, M. 1990. Structure and intonation in spoken language understanding. In Association for Computational Linguistics (ed.) *Proceedings of the 28th Annual Meeting on Associations for Computational Linguistics*. USA: Association for Computational Linguistics, 9-11.

Tan, T. P. 2008. User Manual: Malay Grapheme to Phoneme System. Unpublished document, Universiti Sains Malaysia, Pulau Pinang.

Taylor, P., Black, A., and Caley, R.. 2001. Heterogeneous relation graphs as formalism for representing linguistic information. *Speech Communication*, 33:153–174.

Theune, M., Klabbers, E. A. M., Odjik, J., De Pijper, J., and Krahmer, E. 2001. From data to speech: A general approach. *Natural Language Engineering*, 7:47–86.

Tiun, S., Abdullah, R., and Tang E. K. 2011. Subword Unit Concatenation for Malay Speech Synthesis. *IJCSI International Journal of Computer Science Issues*, 8(5):1694-0814.

Tiun, S., Abdullah, R., and Tang E. K. 2012. Restricted Domain Malay Speech Synthesizer Using Syntax-Prosody Representation. *Journal of Computer Science*. 8(12):1961-1969.

Wightman, C. W. & Ostendorf, M. 1999. Automatic recognition of prosodic phrases. In IEEE Signal Processing Society (ed) *Proceedings of the Acoustics, Speech and Signal Processing (ICASSP 91)*. Washington: IEEE Computer Society, 321-324.

Ye, H. H. 2006. Indexing of Bilingual Knowledge Bank based on the Synchronous SSTC Structure. Master's thesis, University Sains Malaysia, Pulau Pinang.

Zahid, I. & Shah Omar, M. 2006. *Fonetik dan Fonologi*. Kuala Lumpur: PTS Professional.

Sabrina Tiun
Fakulti Teknologi dan Sains Maklumat
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor, Malaysia
sabrinatiun@ftsm.ukm.my
Rosni Abdullah,
Siti Khaotijah Muhammad

Pusat Pengajian Sains Komputer,
Universiti Sains Malaysia,
11800 Pulau Pinang, Malaysia
rosni@cs.usm.my, sitijah@cs.usm.my

Tang Enya Kong
School of Computer Science and Information Technology
Linton University College
Mantin, Negeri Sembilan, Malaysia
enyakong1@gmail.com