Speech Synthesis and Discourse Information

Nick Campbell

[1] National Institute of Information and Communications Technology
[2] ATR Spoken Language Communication Research Laboratory,
nick@nict.go.jp & nick@atr.jp

Abstract

This paper describes some recent work towards a conversational speech synthesis system for use in interactive dialogues between a human and an information system, robot, or speech translation device. The paper describes several response-type utterances that are currently very difficult to implement using traditional speech synthesis methods, and shows how these non-verbal speech sounds function to provide feedback and status-updates in an interactive discourse. The talk will be illustrated with examples of such utterances, which include laughter and grunts as well as common phrases and idiom, showing how their variety can reveal several types of information about the speaker-(i.e., listener) states. The paper proposes a model of information exchange (through speech) whereby this feedback from the listener allows the speaker to efficiently deliver content and to be assured of successful information transmission.

Sinteza govora in diskurzna informacija

Članek opisuje zadnje dosežke pri razvoju pogovornega sintetizatorja govora, ki je namenjen uporabi v interaktivnih dialogih med človekom in informacijskim sistemom, robotom ali govorno-prevajalno napravo. Čanek opisuje več vrst neverbalnih odgovorov, ki jih je z uporabo tradicionalnih postopkov za sintezo govora težko implementirati, in pokaže vlogo teh neverbalnih govornih segmentov pri zagotavaljanju povratne informacije in statusnih osvežitev v interaktivnem diskurzu. Predstavitev bo opremljena s primeri takih neverbalnih govornih segmentov, ki vključujejo smeh in mrmranje, kot tudi pogoste fraze in idiome. Pokazano bo, kako lahko njihova raznolikost razkrije več vrst podatkov o stanju govorca oziroma poslušalca. Članek predlaga model za informacijsko izmenjavo (s pomočjo govora), v katerem poslučalčeva povratna informacija govorcu omogoča, da učinkovito posreduje vsebino in govorcu zagotavlja, da bo informacija uspešno prenesena.

1. Introduction

Speech synthesis has made considerable progress over the past ten years, and some of the recent applications using unit-selection and concatenation of raw waveforms can now only occassionallly be distinguished from natural human speech in terms of voice quality and expressiveness (see for example [1,2]). Their use in many news-reading, announcement, or customer-care applications has become almost transparent, but problems still remain when speech synthesis is to be used in a speech translation environment or when the technology has to replace human speakers in a one-to-one dialogue situation. The expressiveness of a one-to-one conversation is much richer than that of a one-to-many broadcast situation, and many of the differences are signalled using tone-of-voice on utterances that carry little or no propositional content.

2. Data Collection

As part of the JST/CREST 'Expressive Speech Processing' project (ESP), we recorded a series of telephone conversations between ten people who were not initially familiar with each other and who had little or no face-to-face contact during the recording period. They spoke together once a week over the telephone for thirty-minutes each time during a period of three months. The content of the conversations was completely unconstrained. We refer to this as the ESP_C subset of the ESP corpus.

The volunteer speakers were paired so that each conversed with a different combination of partners to maximise the different types of expressiveness in the dialogues without placing the speakers under any requirement to self-monitor their speech or to produce different speaking styles "on-demand".

```
female         male

( cfa efa    cma ema )     (foreign)
   /            \           Group A
 jfa      -     jma
  |              |
 jfb            jmb         Group B
  |              |
 jfc      -     jmc
  |              |          Group C
(fam)          (fam)      (intimate)
```

Figure 1: Showing the form of interactions between the participants. The first letter of the participant identifier indicates the mother-tongue (Japanese/Chinese/English) of the speaker, the second letter indicates the speaker's sex (female or male), and the third letter is the group identifier.

The ten speakers were all recorded in Osaka, Japan, and all conversations were in Japanese. Since the speakers were not familiar with each other initially, the use of the local dialect was not expected and conversations were largely carried out in so-called 'standard' Japanese. Again, no constraints on types of language use were imposed, since the goal of this data collection was to observe the types of speech and the variety of speaking styles that 'normal' people used in different everyday situations.

Four of the ten speakers were non-native; their inclusion was not so that we should have foreign-accented speech data, but rather that we should be able to observe changes in the speech habits of the Japanese native speakers when confronted with linguistically-impaired partners. Two were male, two female, two Chinese, and two English-language mother-tongue speakers. These and the two Japanese who spoke with them formed Group A in our study. Group B is the 'baseline' group, consisting of a male and a female Japanese native speaker who conversed in turn with the each other and with the Japanese native speakers of both sexes from Groups A and C. Group C similarly consisted of a male and a female Japanese native speaker who conversed with each other and with the members of Group B, but who also telephoned their own family members each week and spoke with them for a similar amount of time.

both female	mixed	both male
6425 EFA JFA 7359 JFA EFA		9348 EMA JMA 7433 JMA EMA
8827 CFA JFA 9145 JFA CFA		x (cma jma) 7530 JMA CMA
	9236 EFA JMA 8499 JMA EFA	
	7557 JMA CFA x (cfa-jma)	
	8237 JFA CMA x (cma-jfa)	
	8416 JFA EMA 8560 EMA JFA	
	10068 JFA JMA 7701 JMA JFA	
9069 JFA JFB 9378 JFB JFA		8614 JMA JMB 9465 JMB JMA
8044 JFB JFC 8234 JFC JFB		6983 JMB JMC 7735 JMC JMB
	7686 JFB JMC 7222 JMC JFB	
	10005 JFC JMB 7980 JMB JFC	
13900 JFC Fam		9961 JMC Fam

Table 1: Showing utterance counts and speakers for each recorded conversation. For example, 6425 is the number of utterances spoken by EFA (English female, Group A) to JFA (Japanese female, Group A). There were no sex constraints for speech with family members, but the voice of the remote partner in these cases was not recorded or transcribed. Lower-case shows conversations yet to be transcribed (utterance count shown by an 'x').

The corpus thus allows us to examine the prosodic characteristics and speaking habits of Japanese native speakers when confronted with a range of different partners on the spectrum of familiarity, and to observe changes in their speech as this familiarity changes over time.

Our principal targets for this series of recordings were the six Japanese native speakers (three male and three female) who came to an office building in Osaka once a week to answer the telephone and speak with each partner for a fixed period of thirty-minutes each time. All wore head-mounted close-talking Sennheiser microphones and recordings were taken directly to DAT with a sampling rate of 48kHz. The offices were air-conditioned, but the rooms were large and quiet, and no unwanted noises (or acoustic reflections) were present in the recordings.

```
CFA JFA C01 200.369 0.491 #
CFA JFA C01 200.860 0.808 laugh
CFA JFA C01 201.668 0.869 あと.は
CFA JFA C01 202.537 1.099 変わり.まし.た
CFA JFA C01 203.636 1.868 laugh
CFA JFA C01 205.504 0.670 うん
CFA JFA C01 206.174 0.744 #
CFA JFA C01 206.918 0.917 はい
CFA JFA C01 207.835 2.691 #
CFA JFA C01 210.526 0.602 はい
CFA JFA C01 211.128 2.791 #
CFA JFA C01 213.919 0.749 @S
CFA JFA C01 214.668 2.685 そう.です.結構.も
う.四.年間.なり.まし.た
CFA JFA C01 217.353 0.785 はい
CFA JFA C01 218.138 0.561 #
CFA JFA C01 218.699 0.731 はい
CFA JFA C01 219.430 1.384 #
CFA JFA C01 220.814 1.088 行っ.て.ます
CFA JFA C01 221.902 0.738 #
CFA JFA C01 222.640 0.784 はい
CFA JFA C01 223.424 1.107 #
CFA JFA C01 224.531 1.356 あの.一.歳.です
CFA JFA C01 225.887 0.525 #
CFA JFA C01 226.412 0.600 はい
CFA JFA C01 227.012 2.795 #
CFA JFA C01 229.807 0.443 はい
CFA JFA C01 230.250 0.941 #
```

Figure 2: Transcription was performed by hand, using the Transcriber software package. The first 3 columns identify the speaker, partner, and conversation number. The numbers represent the start time of each utterance in the conversation (in seconds) and its duration. Laughs, non-speech noises, and silences are also transcribed along with the text. Dots in the text represent morphological boundaries as automatically determined by the 'Mecab' software.

The speakers were all mature adults who held part-time jobs with the same company and were paid for their participation in the recordings. They were initially unfamiliar with each other, but the degree of familiarity naturally increased throughout the period of the ten conversations. All have signed consent forms allowing the contents of the recordings to be used for scientific research. The ultimate purpose of the data collection was not made specific to the participants who were only told that their speech would be recorded for use in telecommunications research.

3. Data Analysis

Figure 3 shows the corresponding part of the dialogue segment presented in Figure 2. Here we see the Japanese

```
JFA CFA C01 203.276 1.362 4456 ==> <[ laugh ]>
JFA CFA C01 204.638 0.902 0 ==> <[ @S ]>
JFA CFA C01 205.540 1.927 0 +->  <<あー.,>> .そう.な. << <<ん.です>> .か>>
JFA CFA C01 207.467 0.322 0 ==> <[ @S ]>
JFA CFA C01 207.789 0.401 0 ==> <[ はい ]>
JFA CFA C01 208.190 0.227 0 ==> <[ @S ]>
JFA CFA C01 208.417 1.744 0 ==> <[ あのー ]>
JFA CFA C01 210.976 0.393 814 ==> <[ え ]>
JFA CFA C01 211.369 0.260 0 ==> <[ え ]>
JFA CFA C01 211.629 1.139 0 --> お.,.ご.結婚.を.き.ょ
JFA CFA C01 212.768 0.264 0 ==> <[ え ]>
JFA CFA C01 213.032 1.566 0 --> 何.時.なさっ.た.と.おっしゃい.まし
JFA CFA C01 216.356 0.687 1757 --> 四.年.目
JFA CFA C01 217.043 0.301 0 ==> <[ @S ]>
JFA CFA C01 217.344 1.498 0 +-> あ.,. <<そう.です>> .か
JFA CFA C01 218.842 0.422 0 ==> <[ @S ]>
JFA CFA C01 219.264 0.241 0 ==> <[ え ]>
JFA CFA C01 219.505 1.193 0 --> お子さん.は
JFA CFA C01 221.686 0.283 987 --> X
JFA CFA C01 221.969 0.819 0 --> あ.,.いらっしゃる
JFA CFA C01 223.180 0.360 392 ==> <[ あ ]>
JFA CFA C01 223.540 1.248 0 --> お.幾つ.です.か
JFA CFA C01 225.571 0.749 783 --> 一.才
JFA CFA C01 226.320 0.347 0 ==> <[ @S ]>
JFA CFA C01 226.667 1.235 0 --> あっ.そう.,.じゃ
JFA CFA C01 227.902 1.891 0 +-> こういう. <<とき.は>> .どういう.風.に
JFA CFA C01 229.793 1.494 0 +-> お子さん.は.さ.れ.てる. << <<ん.です>> .か>>
JFA CFA C01 231.287 0.746 0 ==> <[ @S ]>
JFA CFA C01 232.033 0.798 0 --> お家
JFA CFA C01 234.539 0.424 1707 ==> <[ あ ]>
```

Figure 3: The corresponding part of the dialogue shown in Figure 2, but after processing to identify repeated patterns. Frequent utterances $(n >= 100)$ are shown in $< [square] >$ brackets, and frequent segments $(N >= 100)$ within longer utterances are shown in $<< angle >>$ brackets, which may be embedded. Also shown here in column 6 are the delays (in milliseconds) between succeedding utterances.

speaker's utterances and can combine them with those of her Chinese partner to reproduce the conversation segment. Some potentially ambiguous utterances can thereby be disambiguated by use of the textual content of the surrounding utterances, but a large number remain functionally indeterminate from the transcription alone. They are not at all ambiguous when listening to the speech, and carry a considerable amount of discourse information.

JFA:	CFA	CMA	EFA	EMA	JFB	JMA
a,a–	143	145	88	89	138	170
ano	224	277	221	176	209	266
demo	41	24	31	17	89	134
e–	48	51	37	25	74	94
hai	2932	2234	2181	3239	72	33
un,un	1029	546	585	1190	909	1037

Table 2: Counts for some frequently-repeated simple utterances from one speaker to six partners. The table illustrates differences in usage strategies for these utterances.

The text in Figure 3 has been further annotated by a computer program to show which utterances are unique (and therefore presumably convey more propositional content) and to mark those which are subject to frequent repetition (and hence portray affect or discourse-control information). Two types of repetition have been marked; (a) whole phrases, and (b) phrasal chunks that form part of a larger, possibly unique, utterance but which are frequently repeated anyway. The chunks were determined by use of the pds 'mecab' software [3] for morphological decomposition, in conjunction with 'yamcha' [4] for regrouping of the fine morphological segments.

The current setting of the pattern recognition program, arbitrarily taking more than 99 repeats throughout the corpus as the minimum threshold for bracketing, yields 74,324 untouched utterances, 72,942 marked as repeated phrases, and 49,136 utterances including repeated phrasal segments.

Taking some of the frequent repetitions from one of the corpus speakers as an example, we notice different strategies of usage according to differences in partner. This speaker (JFA) makes considerable use of "a", "ano", "hai", and "un", but not equally with all partners (see Table 3). For example, when speaking with foreigners, she uses "hai" (はい= yes(perhaps?!)) frequently, but significantly less so when speaking with Japanese partners. She uses "demo" (でも=but) much more frequently with Japanese partners, and "a" much less when conversing (in Japanese) with the

10073	うん	467	ズーー	228	ううん	134	へー.ーー
9692	@S	455	スー	227	えっ	134	はい.はい.はい.はい
8607	はい	450	んーー	226	へ.ーー	134	そう.です
4216	laugh	446	うーーーん	226	ハハハ	133	@E
3487	うーん	396	ねー	225	う.んー	133	あ.そう.な.ん.です.か
2906	ええ	395	あ.あー	200	そうですね	130	そう.な.ん.です.か
1702	はーい	393	はい.はい.はい	199	ほ.ーー	129	は.ー
1573	うーーん	387	あー.はい	193	ハー	129	い
1348	ズー	372	ねえ	192	その	127	ほ.ー
1139	ふん	369	ふーーん	190	え.えー	125	ハハハハハ
1098	あのー	369	だから	188	あ.あーー	119	はい.はい
1084	あっ	368	あー.ん	187	ね	119	は.ーーー
981	はあい	366	ああ	180	ん.はい	114	ハハ
942	あの	345	あの.ーー	180	あの.ーーー	113	は
941	ふーん	337	なんか	173	ん.ん	113	で.ー
910	そう	335	え	172	アハハハ	113	て
749	えー	311	でも	168	はい.ー	112	は.あー
714	あーー	305	スーー	164	う.うーん	110	フフフ
701	あ	274	うん.うん.うん	161	は.ーー	110	そのー
630	あーーー	266	ハハハハ	160	@K	110	もう
613	あ.はい	266	て.ー	159	そう.です.ねー	109	ふーーーん
592	うん.うん	266	え.ーー	151	あーーーー	108	はあ.ーー
555	あー	258	で	143	だから.ー	106	そうですね.え
500	んー	248	う	139	アハハハハ	105	んー.ん
469	ん	242	へー	137	そう.そう.そう	104	いや

Table 3: The hundred most frequent single utterances in the ESP_C corpus. The numbers indicate the count of each word or phrase when it occurs as a single utterance in the transcriptions. Since duration is usually considered as distinctive in Japanese, the lengthening (an extra mora beat is indicated by a dash) may be significant. Note the highly repetitive nature of many of these utterances, very few of which can be found in any standard dictionary of Japanese. Note that these few samples alone account for more than a third $(n = 72,685)$ of the 200,000 utterances in the corpus. Less then half $(n = 92,541)$ of the utterances were unique.

English-native-speaker partners.

Such differences may reflect interpersonal relationships, personal characteristics, or cultural peculiarities, but perhaps more interesting to us here (with speech synthesis in mind) is the variety of pronunciation within each utterance type, reflecting the speaker's interest, state-of-mind, and type of participation in the discourse.

4. Ambiguous Utterances — — A Challenge for Synthesis

It is a central tenet of this paper that these repeated segments can be used to carry affect-related and interpersonal information by variation in such acoustic characteristics as tone-of-voice, spectral tilt, pitch range and excursion, speaking rate, phonatory setting, etc. By being frequent and repetitive, they allow the listener (even one not yet familiar with the speaker's traits) to make comparative judgements about the speaker's emotional and affective states and stances and to interpret subtle nuances in the speech by means of the prosodic cues hereby revealed [5].

In speech synthesis, a given text sequence is rendered into speech with a given prosodic pattern, usually predicted from part-of-speech information in conjunction with the position of the words in the phrase and sentence. Here, however, we have whole phrases that consist of a single word (itself often of doubtful or indeterminate part-of-speech status) whose prosody is dependent upon the speaker's affective states and discourse intentions; there is currently no way of easily specifying these higher-level constraints in a synthesiser apparatus.

"a,a–"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	125	181	266	232	234	241
f0m	201	214	220	192	206	198
pwr	28	29	29	28	31	31
pwm	38	39	36	35	42	41
"un,un"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	154	152	182	181	161	141
f0m	172	175	162	145	198	174
pwr	28	29	27	26	29	27
pwm	37	40	36	35	42	39
"ano"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	106	113	161	154	169	155
f0m	131	136	142	133	156	149
pwr	27	28	28	27	31	29
pwm	38	40	37	36	42	39

Table 4: F0 range (f0r) and average (f0m) values in Hz and Power range (pwr) and average (pwm) values in dB for three sample utterances from speaker JFA according to differences in conversational partner (see Figure 4).

Tables 3 and 4, and Figure 4 illustrate some diffferences in pitch range (i.e, the amount of variation in the funda-

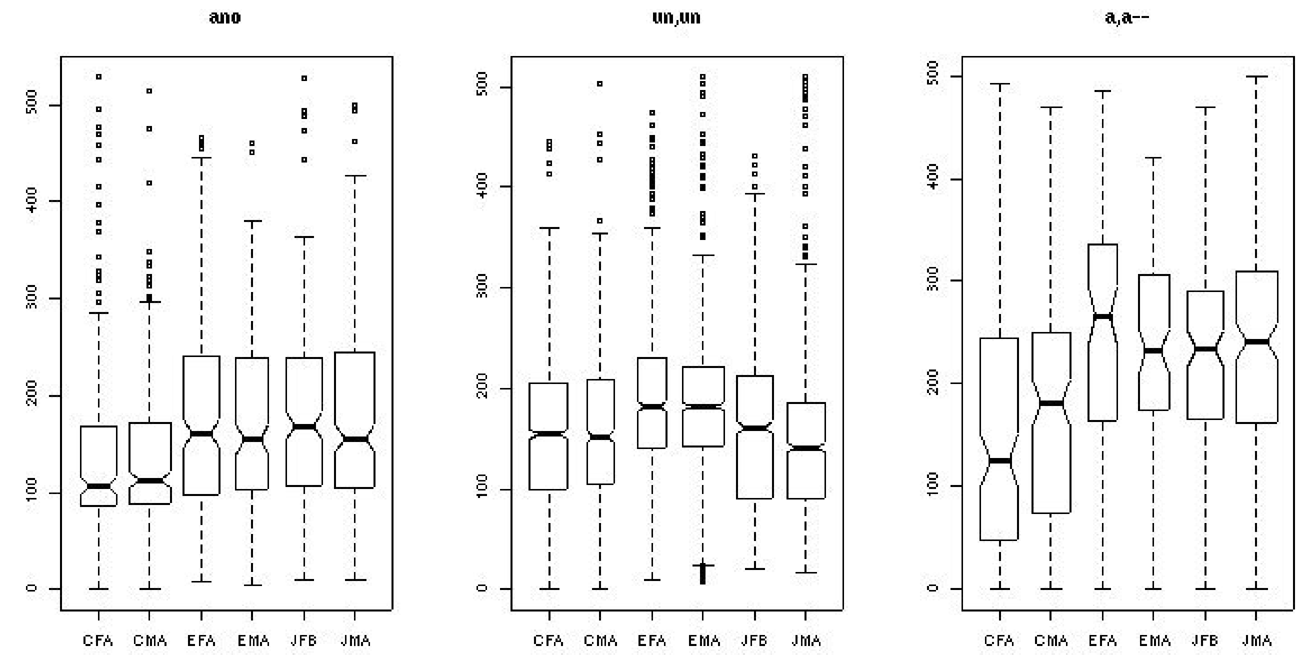

Figure 4: Plots of Pitch Range (amount of variation in the fundamental frequency of the voice) for three utterances from speaker JFA when conversing with six different partners. The width of the boxes is proportional to the number of tokens. Differences are significant at the 5% level if the notches do not overlap. The vertical axis shows pitch range in Hz.

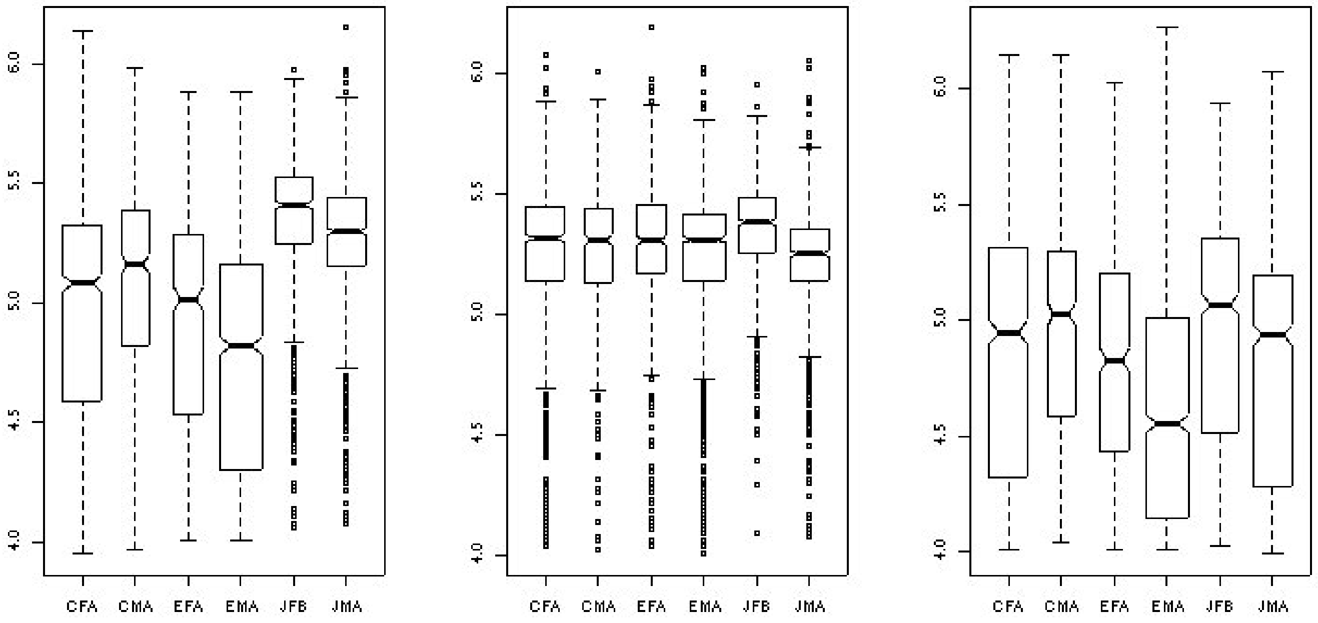

Figure 5: Fundamental frequency contours differ according to the listener. The left-hand plot shows average f0 values for the initial third of the utterance, the middle plot for the middle third, and the right-hand plot shows average f0 values for the final third of the utterance. Plots show‘contours’ for “un,un”. We can see that Japanese partners evoke a high initial contour, and English-native-speakers a lower fall at the end, though all contours appear to pass through the same high range of values mid-utterance.

mental frequency of the voice throughout the utterance) and voice energy (signal power in decibels) for three representative but randomly-selected sample utterances from speaker JFA's conversations with six different partners.

The data show that the speaker's basic acoustic settings and amount of physical energy used in each utterance vary not just by utterance, as would be expected, but also by listener (and presumably according to the content of the conversations). Figure 5 takes a subset of this data (f0 contours for the utterance "un,un") and plots a representation of the 'shape' of each utterance by showing averaged f0 values for each progressive third of the utterance. Again we see considerabe variation, but that the variation between contours for different types of conversation partner is greater than that between utterances within a given set of conversations.

We can see that Japanese partners evoke a high initial contour, and English-native-speakers a lower fall at the end, though all countours appear to pass through the same high range of values mid-utterance. The fact that these differences appear more related to partner than to local contextual differences implies that a higher-level of prosodic processing is taking place; i.e., that a level of social interaction is influencing the prosodic contour just as the linguistic relations influence it a lower more independent level.

5. Conclusion

This paper has presented some data from the ESP_C corpus of conversational dialogues, and has shown that there is considerable prosodic variation on what are seemingly very simple but also very frequent utterances. This variation may indicate the speaker's relationship with the listener, since it seems to vary more between conversational partners than between different utterances.

From a speech synthesis standpoint, this data presents problems for current systems which use one standard set of rules for predicting all prosodic characteristics. These rules currently make no allowance for difference in the relationship with the listener (or conversational partner) but for interactive speech synthesis systems where a computer is generating speech on behalf of one partner, such as in a speech translation system, such information must be mapped, processed, and included in the prosody control rules.

Acknowledgement

This work was supported by the Japan Science & Technology Agency (JST/CREST) and the National Institute of Information & Communications Technology. The author also wishes to thank the management of ATR SLC for support.

6. References

Beutnagel, M., Conkie, A., Schroeter, J., Stylianou, Y., and Syrdal, A., "The AT&T Next-Gen TTS System", in Proc TC-STAR Workshop on Speech-to-Speech Translation, Barcelona, Spain, 2006.

Campbell, N., "Conversational Speech Synthesis and the Need for Some Laughter", in *IEEE Transactions on Audio, Speech, and Language Processing, Vol 14, No.4*, July 2006.

Mecab: http://mecab.sourceforge.jp/

Yamcha: http://www.chasen.org/ taku/software/yamcha/

Campbell, N., "Getting to the heart of the matter; speech as expression of affect rather than just text or language", pp 109-118, *Language Resources & Evaluation Vol 39, No 1*, Springer, 2005.