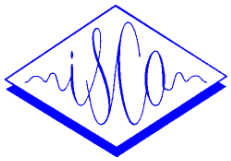

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

TAXONOMY OF SPONTANEOUS SPEECH PHENOMENA IN MANDARIN CONVERSATION

Shu-Chuan Tseng

Institute of Linguistics, Academia Sinica
tsengsc@gate.sinica.edu.tw

ABSTRACT

Spontaneous speech raises a number of research issues which cannot be observed in other types of speech data. Disfluent speech, ill-formed sequences and particular pronunciation variations mark the most important facet of spontaneous speech. The goal of this paper is to provide a taxonomy scheme of spontaneous speech phenomena, which offers the necessary basis for research works and applications dealing with spontaneous speech. The proposed taxonomy focuses on the phonetic and structural peculiarity and subsequently, the adequacy of this taxonomy is evaluated by empirically examining a corpus of annotated Mandarin data.

1. INTRODUCTION

Most of the established speech recognition and parsing systems can successfully process well-formed and well-spoken utterances, for instance clearly read speech. But for the rest of "ill-formed" and "not properly spoken" utterances found in spontaneous conversation, no satisfying solutions have been found yet. All these fragmentary, incomplete or even regarded as incorrect speech stretches are by no means marginal phenomena. A number of empirical studies on repairs, pauses and disfluencies have been done by Levelt [1], Hirschberg and Litman [2], Nakatani and Hirschberg [3] and the others. However, it is hardly possible to define the domain of spontaneous speech phenomena consistently across different corpora and the annotation scheme used for each corpus varies to a great extent such as in the Map Task Corpus [4], the Switchboard Corpus [5], the TRAINS Corpus [6] and the BAUFIX Corpus [7].

This paper systematises types of spontaneous speech sequences occurred in Mandarin conversation and proposes a taxonomy scheme for producing well-prepared spontaneous speech database. Mandarin examples are written in Pinyin and the numbers following each syllable are lexical tones. 1, 2, 3, 4, and 5 represent high flat, rising, contour, falling tones and the neutral tone.

Part of the research study presented in this paper was funded by the NSC (91-2411-H-001 -043) and the Ministry of Education.

2. TAXONOMY OF SPONTANEOUS SPEECH SEQUENCES

Spoken sequences in spontaneous speech whose linguistic components do not follow conventional regulations or cannot be judged exactly right or false by linguistic norms are considered in this paper. The taxonomy deals with the phonetic and the structural features. We do not separate all the phenomena into phonetic and structural groups, but categorise them by their linguistic characteristics and functions in conversation. Examples are for instance pronunciation deviated from the standard phonological regularities, prosodic discontinuity in pitch and constructions not in agreement with the standard grammatical rules such as incomplete sentences and speech repairs. Therefore, four groups are proposed: 1) disfluency, 2) socio-linguistic phenomena, 3) particular vocalisation and 4) unintelligible and non-speech sounds. In the following, we only give concise explanations due to the lack of space, detailed operational definitions of each tag can be found in [8].

2.1. Disfluency

2.1.1. Prosodic Disfluency

Silence occurs, when none of the speakers utters anything to fill the silent gap within the course of conversation. Sometimes, it is so short, that it hardly bothers the conversation participants. But sometimes it can last for a few seconds and often embarrasses all participants. Another sort of prosodic disfluency is *pause*. By definition, pause only occurs within a certain speaker's turn. In spontaneous speech, pausing does not necessarily follow the framework of syntactic phrasing. Depending on the communication situation and the intention of the speaker, the pausing can be located anywhere within utterances irrespective of the standard syntax. *Short breaks* in real situations are more often found than pauses and they are shorter than pauses. Short breaks sometimes are produced due to physical needs such as breathing or clearing throat and are accompanied by acoustically measurable signals. In conversation, speakers may hesitate or *stutter*, when they have problems in finding the

correct words. Hesitation is usually realized as pause or silence. In Mandarin, each syllable is represented by a character and means something, so each uttered syllable may be a possible candidate for a word. For this reason, we suggest that in the case of Mandarin the stuttering has to be handled separately.

2.1.2. Repair

We propose a system of repairs, divided into six categories *restart*, *repetition*, *overt repair*, *editing term*, *error*, and *word fragment*. Repetitions in Mandarin are in a lot of cases perfectly legal syntactic constructions to put emphasis on particular components or to express subtle semantic nuance. But repetitions in this context are disfluent repetitions, which cannot be explained or justified by Mandarin grammatical rules. *Repetitions* are defined as fully repeated word sequences, where partial repetitions are grouped into the category *restart*. *Overt repairs* are repairs with both the reparandum and the reparans. Only disfluent sequences in which we can clearly identify what is to be corrected and what is the correction are regarded as overt repairs. In the sequence "shi4 jin4kou3 EN chu1kou3 ma1" ("shi4" means the verb BE, "jin4kou3" import, "chu1kou3" export, "EN" is a discourse particle and "ma1" is a grammatical particle for interrogative sentences), "jin4kou3 is the reparandum and "chu1kou3" is the reparans. Also found in this example, "EN" is an *editing term*, often used to bridge the gap between the reparandum and the reparans. Words not completely uttered are marked with the category *word fragment*. In Mandarin, phonetic errors are hardly identifiable without knowing the subsequent syllables. Thus, the category *error* is only used for lexical or syntactic errors. False combination in compound words, idioms, and false classifiers for nouns belong to this category, too.

2.1.3. Syntactic Disfluency

Mandarin is a language with extremely limited restrictions on constituting grammatically correct sentences. Ellipsis of subjects and objects is completely legal. No morphological marking is needed to identify the well-formedness of sentences. Hence, we introduce three categories to group what we call "syntactic disfluency": *inappropriate*, *abridged*, and *interrupted* utterances. When an utterance sounds inappropriate and there is more than one way (to add, to delete or to replace something) to make the utterance sound correct, the utterance is marked as *inappropriate usage*. If an utterance is incomplete from the syntactic point of view, it is defined as *an abridged utterance*. Incomplete utterances resulting from other speakers are defined as *interrupted* utterances. We invent these two categories to preserve different kinds of possibilities to relate the subsequent utterance to the abridged and interrupted utterances. Abridged utterances may have a stronger relationship to the next utterance than an interrupted utterance that simply stops at the time point of interruption.

2.1.4. Discourse Particles and Markers

In our notation, we consistently use capital letters to transcribe discourse markers, particles and other non-lexicalised items to distinguish them from words having the same pronunciation and similar meaning. Some of the discourse particles used in Chinese can be exactly mapped to characters under the condition that a consensus on a unified semantic use is shared among the native speakers. However, for some *discourse particles*, it is difficult to choose the exactly matched character to transcribe. Besides, there exist no characters for discourse particles originated from dialects, either.

Discourse markers are words whose use in spontaneous speech deviates from their original semantic meaning and syntactic role and instead, their pragmatic function increases. For instance, "na4" originally means *that* being a determiner or *that* being a pronoun and "na4ge5" means *that* + classifier. When used in discourse, NA often appears in the utterance-initial position, used for signalling the speaker intention to begin a new turn. Clearly different from NA, NAGE is usually found in the mid-utterance position, also signalling the beginning of a piece of new information. However, in our data NAGE is found located before many syntactic categories and has the function of more than a definite article [9].

2.2. Socio-linguistic Phenomena

Socio-linguistic phenomena take into account features or sequences which indicate socio-linguistic characteristics. English, Japanese and dialects such as Min-Nan are frequently used in everyday life in Taiwan. When the currently used language is changed to a foreign language or a dialect, it is marked with *code switching*. *Dialect-influenced pronunciation* is a variation of pronunciation influenced by dialects. Newly created fashion words which are mostly invented and used by young people are marked with *new word*.

2.3. Particular Vocalisation

Particular vocalisation covers special pronunciation variations in spontaneous speech. *Lengthening* marks a lengthened speech sound. Sounds which are not nasal but produced in a nasalised manner are annotated with *nasalised*. A contraction of syllables due to deletion or assimilation of neighbouring speech sounds is marked with *syllable contraction*. *Assimilation* marks the assimilated adjacent speech sounds. All other deviated pronunciation variations which

cannot be specifically grouped into any category above are marked with *inappropriate pronunciation*.

2.4. Unintelligible and Non-speech Sounds

Non-speech sounds include all recognisable verbal but non-speech sounds for instance laughing and coughing etc. as well as all non-verbal sounds such as noises. All other unintelligible speech or non-speech sounds are marked with *unintelligible sounds*.

3. EXPERIMENTAL RESULTS

This section presents empirical results of annotating Mandarin spontaneous speech by applying the above taxonomy.

3.1. Data

Mandarin Conversational Dialogue Corpus was collected from 2000 to 2001 at the Institute of Linguistics in Academia Sinica. It consists of 30 digitized conversational dialogues of a total length of 27 hours. 60 subjects were randomly chosen from the capital city, Taipei. Eight conversations spoken by nine female and seven male speakers were annotated by adopting the taxonomy scheme above. The dialogues are represented by d-01, d-02, d-03, d-04, d-05, d-06, d-07 and d-08. To undertake this experiment, we used the interface "TransList" [10] to transcribe the conversations in Chinese characters and in Pinyin, to insert the annotation tags and to convert the horizontally arranged transcripts to a character-based and vertically presented database in Access format.

3.2. Annotated Results

53,225 annotation tags were used to annotate totally 140,579 transcribed characters by five human transcribers. Table 1. illustrates the results in terms of the proportion of each group and the two most frequently used tags in brackets. The gender of the speakers and the total number of the produced characters in each conversation are also given in the table. Disfluency constitutes the largest group, followed by particular phonetic variations. Among disfluency, particles and short breaks are dominantly produced in spontaneous Mandarin across all eight conversations. Particles make up 15.34% of the totally annotated tags; short breaks 8.98%. This result clearly shows the importance of the small things such as particles and short breaks in speech applications concerned with spontaneous speech. Short breaks are related to pausing in speech, therefore also influences the identification of constituent boundaries. Similarly, other phonetic variations are also often found in our data such as syllable contractions, lengthening, and inappropriate pronunciation. Missed or mismatched syllables may result in serious recognition problem in Mandarin. Inappropriate pronunciation makes up 10.67%. For these sequences, the human transcribers cannot exactly tell what kinds of phonological deviations they are. For a recognition system dealing with spontaneous Mandarin, further investigations into this group of production variations are necessary. We surprisingly identified a relatively large number of lengthening in the data, 4.99%. Pause makes up about the same percentage, 4.2%. It seems that speakers also use lengthening, when they hesitate. Used with pauses together, a lengthened sound is often much longer than a pause. Code switching identified in the data includes the use of English, Japanese, and Min-Nan, which is a dialect dominantly spoken in Taiwan. The use of non-Mandarin languages is marked in about 0.5% of the overall data.

speakers charac.	**disfluency**	**socio-linguistic phenomena**	**particular vocalisation**	**unintell. non-speech s.**
d-01	**34.45%**	**1.13%**	**48.37%**	**16.05%**
fem.-mal.	(particle)	(English)	(syll. con.)	(unrecog. s.)
18833	(short break)	(Japanese)	(inapp. pro.)	(inhale)
d-02	**44.83%**	**0.68%**	**34.47%**	**20.03%**
fem.-mal.	(short break)	(English)	(syll. con.)	(unrecog. s.)
17515	(particle)	(Min-Nan)	(length.)	(inhale)
d-03	**50.74%**	**0.07%**	**33.69%**	**15.49%**
fem.-fem.	(particle)	(Min-Nan)	(syll. con.)	(unrecog. s.)
14632	(short break)		(length.)	(inhale)
d-04	**44.07%**	**0.43%**	**47.26%**	**8.24%**
fem.-mal.	(particle)	(English)	(inapp. pro.)	(unrecog. s.)
23817	(short break)	(Min-Nan)	(syll. con.)	(inhale)
d-05	**33.52%**	**0.67%**	**58.42%**	**7.39%**
fem.-fem.	(particle)	(Min-Nan)	(syll. con.)	(unrecog. s.)
20689	(short break)	(English)	(inapp. pro.)	(inhale)
d-06	**34.98%**	**0.14%**	**56.35%**	**8.53%**
mal.-mal.	(particle)	(English)	(syll. con.)	(unrecog. s.)
16833	(short break)	(Min-Nan)	(inapp. pro.)	(inhale)
d-07	**51.29%**	**8.83%**	**31.59%**	**8.29%**
fem.-mal.	(particle)	(Min-Nan)	(syll. con.)	(unrecog. s.)
15806	(short break)	(English)	(inapp. pro.)	(inhale)
d-08	**30.36%**	**0.25%**	**43.42%**	**25.97%**
fem.-mal.	(particle)	(English)	(syll. con.)	(unrecog. s.)
12454	(short break)	(Min-Nan)	(inapp. pro.)	(inhale)

Table 1. Annotated Results

3.3. Adequacy of the Proposed Taxonomy

Figure 1. illustrates the distribution of the used tags in all eight conversations. Across the eight conversations and the five transcribers, the distribution shows a clear pattern of regularity and similarity. Short break, particle, syllable contraction, repetition, abridged utterance, lengthened sound

and inappropriate pronunciation are the most frequently perceived and marked spontaneous speech phenomena. Phonetic phenomena such as short break, syllable contraction, lengthening and inappropriate pronunciation need further experimental studies into their acoustic features, where pragmatic and syntactic analyses are necessary for investigating the structural peculiarity of spontaneous speech such as particles, repetitions and abridged utterances.

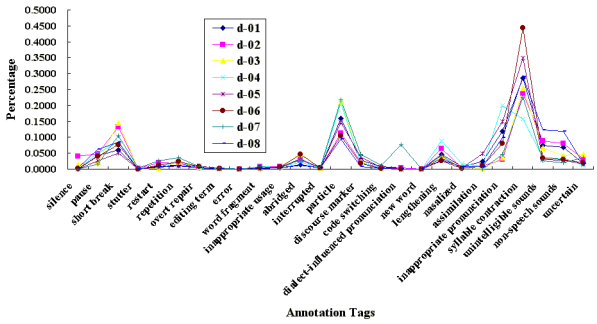

Fig. 1. Annotated Results in Terms of Conversations

To statistically test the adequacy of the proposed taxonomy, we calculated the correlation of the annotated results in terms of individual conversations. We obtained the following similarity matrix of correlation.

	d-01	d-02	d-03	d-04	d-05	d-06	d-07	d-08
d-01		.894	.894	.841	.969	.925	.858	.946
d-02			.930	**.677**	.820	.871	.823	.924
d-03				**.769**	.841	.842	.939	.832
d-04					.801	**.646**	**.777**	**.701**
d-05						.954	.827	.882
d-06							.800	.898
d-07								**.732**
d-08								

Table 2. Correlation Matrix

In the similarity matrix, all eight conversations show a high and positive correlation with each other. Only the correlation coefficients calculated for the conversation d-04 are relatively small in relation to d-02, d-03, d-06, d-07 and d-08. This may have to do with the nature of the particular conversation or the particular transcriber. Nevertheless, the taxonomy proposed in this paper, as a whole, was consistently applied by human transcribers to annotate spontaneous speech phenomena.

4. CONCLUSION

To sum up, the annotation system proposed in this paper is applicable for annotating Mandarin conversational dialogues. The distribution of annotated phenomena across all eight conversations shows similar patterns. This indicates the fact that in the proposed system the phenomena are well classified and well defined. Furthermore, the most frequently produced sequences in each classification group are all represented by a similar set of annotation tags. This supports the notion that spontaneous speech phenomena do not expand onto an infinite set, but regularly fall into a certain range of linguistic definition of phenomena. The proposed taxonomy is by no means complete yet. But what was argued in this paper is that prior considerations of the linguistic features of spontaneous speech are necessary and useful before the actual design and implementation of applications related to spontaneous speech.

5. REFERENCES

[1] W. J. Levelt, "Monitoring and Self-Repair in Speech," *Cognition*, vol. 14, pp. 41–104, 1983.

[2] J. Hirschberg and D. Litman, "Empirical Studies on the Disambiguation of Cue Phrases," *Computational Linguistics*, vol. 19, pp. 501–530, 1993.

[3] C. Nakatani and J. Hirschberg, "A Corpus-Based Study of Repair Cues in Spontaneous Speech," *Journal of the Acoustical Society of America*, vol. 95, pp. 1603–1616, 1994.

[4] J. Carletta, R. Caley, and S. Isard, "A Collection of Self-Repairs from the Map Task Corpus," Tech. Rep., University of Edinburgh, 1993.

[5] E. Shriberg, "Disfluencies in SWITCHBOARD," in *Proc. of the International Conference on Spoken Language Processing*, Philadephia, 1996.

[6] Peter Heeman and James Allen, "Speech Repairs, Intonational Phrases and Discourse Markers: Modelling Speakers' Utterances in Spoken Dialogue," *Computational Linguistics*, vol. 25, no. 4, pp. 527–571, 1999.

[7] G. Sagerer, H. Eikmeyer, and G. Rickheit, ""Wir bauen jetzt ein Flugzeug": Konstruieren im Dialog. Arbeitsmaterialien," Tech. Rep., SFB 360 "Situierte Künstliche Kommunikatoren". Universität Bielefeld, 1994.

[8] Shu-Chuan Tseng and Yi-Fen Liu, "Annotation Manual of Mandarin Conversational Dialogue Corpus, 02-01," Tech. Rep., CKIP, Academia Sinica, 2002.

[9] Shuanfan Huang, "The Emergence of a Grammatical Category Definite Article in Spoken Chinese," *Journal of Pragmatics*, vol. 31, no. 1, pp. 77–94, 1999.

[10] Shu-Chuan Tseng and Yi-Fen Liu, "Mandarin Conversational Dialogue Corpus. MCDC 2001-01," Tech. Rep., Institute of Linguistics, Academia Sinica, 2001.