# Generating Fillers based on Dialog Act Pairs for Smooth Turn-Taking by Humanoid Robot

Ryosuke Nakanishi, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara

(Kyoto University, Japan)

# Background

## ◆ Current Spoken Dialog Systems (SDS)

Smartphone


Smart Speaker

Natural
interaction

Humanoid robot

## ◆ Turn-taking Protocol

- push-to-talk or magic word
- GUI or LED

→ natural behaviors
using fillers & backchannels

# Related Works

- Prediction of fillers for language modeling of ASR [Akita10]
  - Fillers are to be removed
- Generation of fillers in speech synthesis [Anderson10, Sundaram14]
  - Fillers have different prosody
- Prediction & generation of backchannels [Kawahara16]…
  - For attentive listening
- Generation of fillers in dialog [Shiwa08, Skantze14]
  - Only simple forms

# Roles of Fillers

- Signals thinking & hesitation
- Improves comprehension
  - Provide time for comprehension [Watanabe09]

- Attracts attention & improves politeness
  - Mitigate abrupt speaking

- Smooth turn-taking
  - Hold the current turn, or Take a turn

- Improve naturalness?

# Outline (Research Questions) of this Study

**Investigate effect of fillers for smooth turn-taking in SDS**

1. **Analysis of filler occurrence and forms in dialog act (DA) pairs**
   - Do fillers occur more often when turn-keep/switch is ambiguous?
   - Are there specific types of fillers depending on DA pairs?

2. **Prediction of fillers**
   - How well can we predict filler occurrence and forms?
   - Effect of DA pairs?
   - Effective features?

3. **Generation of fillers**
   - Does it help avoid speech collisions?
   - Naturalness and likability

# Corpus of Dialog with ERICA in WOZ setting


Remotely operated


- **#sessions (subjects):** 39 (male: 16名, female: 23)
- **Duration of dialog:** 10 min.
- **Roles in dialog:** subjects = visitor to lab, ERICA = secretary

# Annotations

◆ **Fillers** [Koiso06]
- when turns are held or taken
- Flat pitch

あ(ー), え(ー), え(ー)(っ)と, う(ー)ん, なんか(ー),
ま(ー), あの(ー), その(ー)

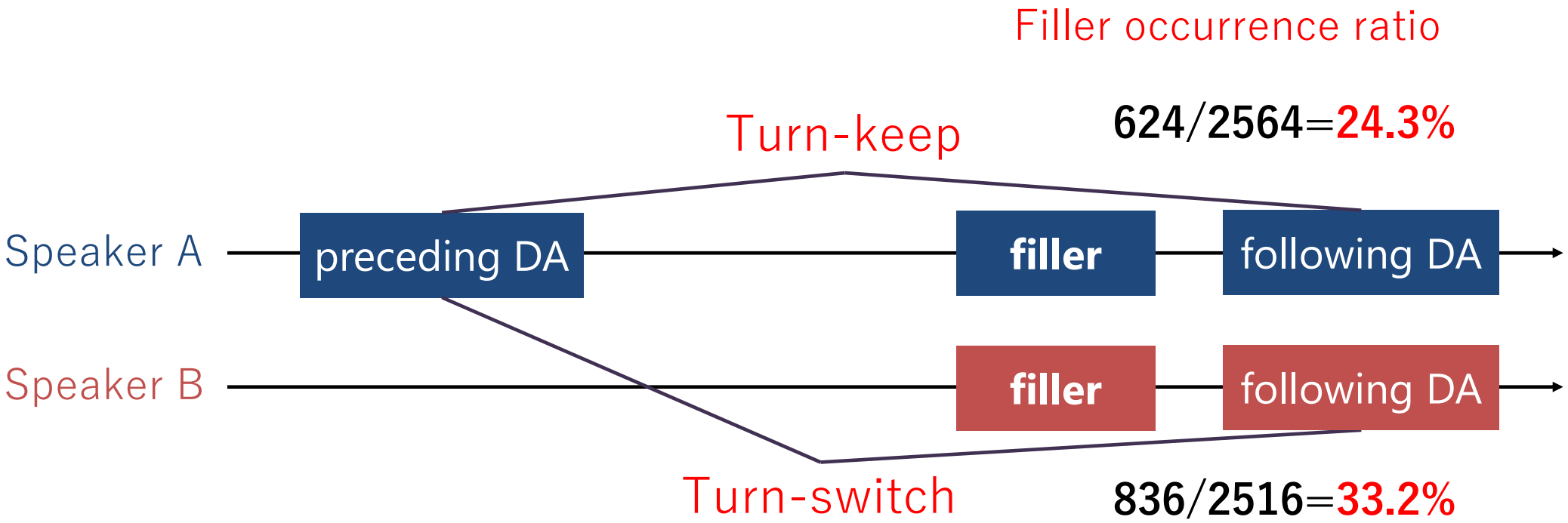
◆ **Backchannels**
- when turns are not taken

◆ **Dialog Act: DA** [Bunt10]

– **Question (Q):** information-seeking
– **Statement (S):** inform/offer/promise/request/instru
– **Response (R):** answer/accept offer/decline offer
– **Others (O):** greeting/apology..

| | operator | subjects |
|---|---|---|
| Q | 758 | 267 |
| S | 1064 | 1687 |
| R | 779 | 477 |
| O | 706 | 703 |

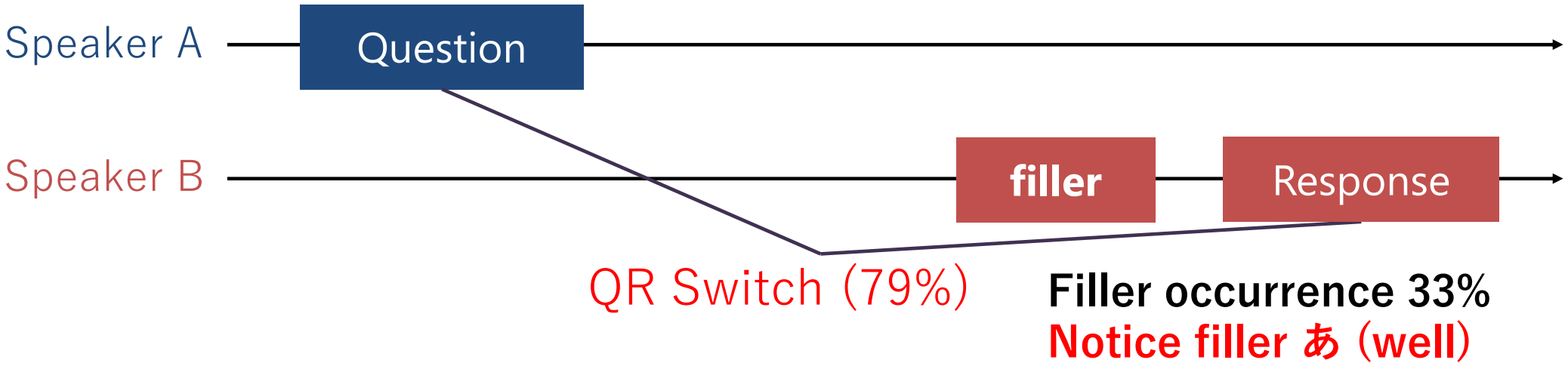
# DA Pair and Filler Occurrence




**Total 1460/5080=28.7%**

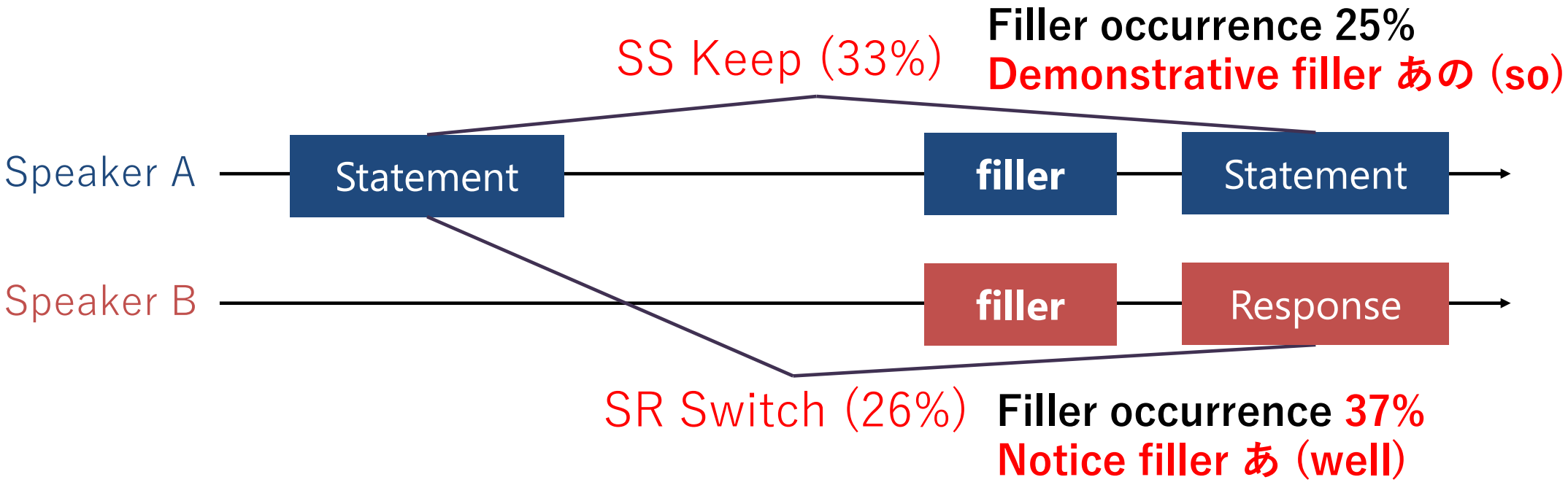
**Fillers occur more frequently in turn-switch than in turn-keep**

# Typical DA Pair and Fillers (QR Switch)



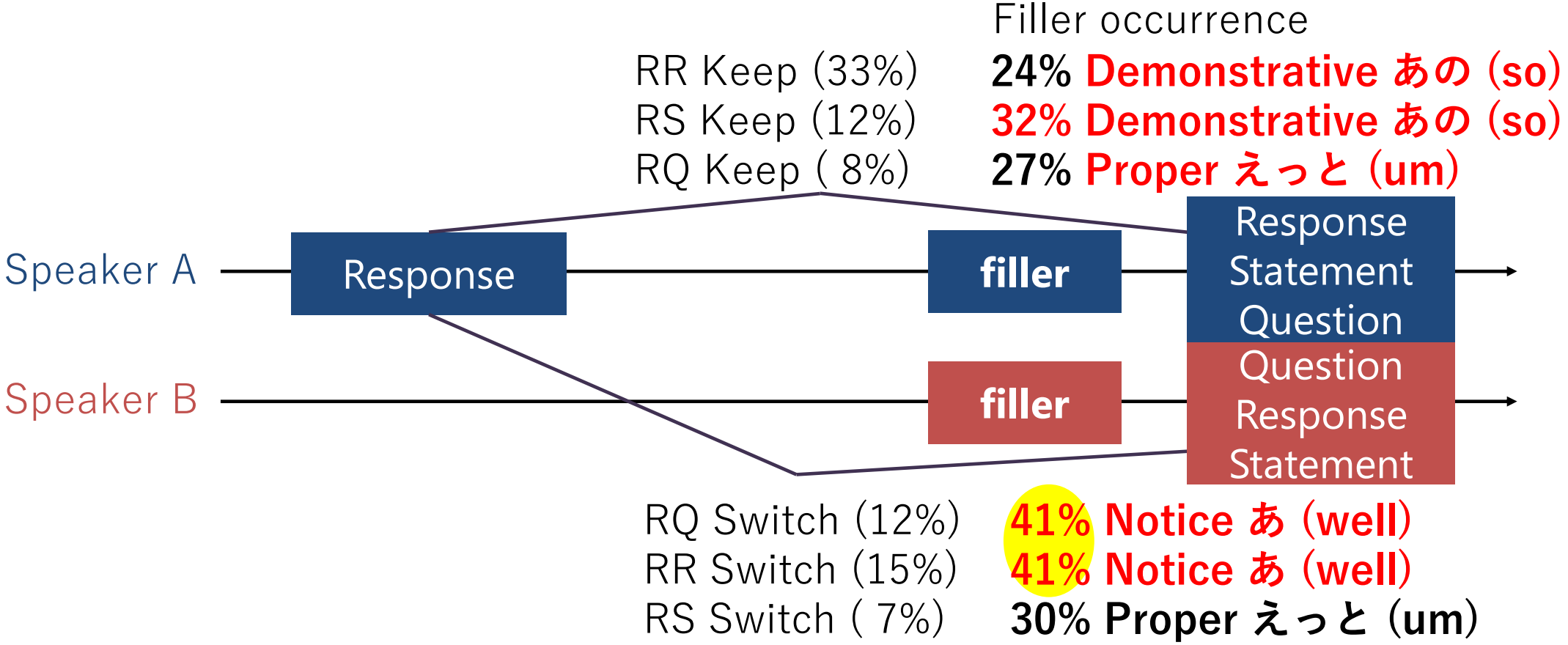

- After Question, turn-switch and Response are expected.
- Notice fillers (ack.) are typically used.

# Typical DA Pair and Fillers (SS Keep/SR Switch)




- There is ambiguity in turn-switch/keep after Statement.
- Typical DA and filler pattern exist for each interlocutor.

# Typical DA Pair and Fillers (SS Keep/SR Switch)




- **After Response, much ambiguity both in turn-switch/keep and following DA**
- **Fillers occur most frequently in turn-switch**

# Tendencies of Filler Occurrence and Forms

- **Fillers occur more frequently in turn-switch than in turn-keep**
- **Fillers occur most frequently in turn-switch after Response (R), when the following DA is ambiguous**

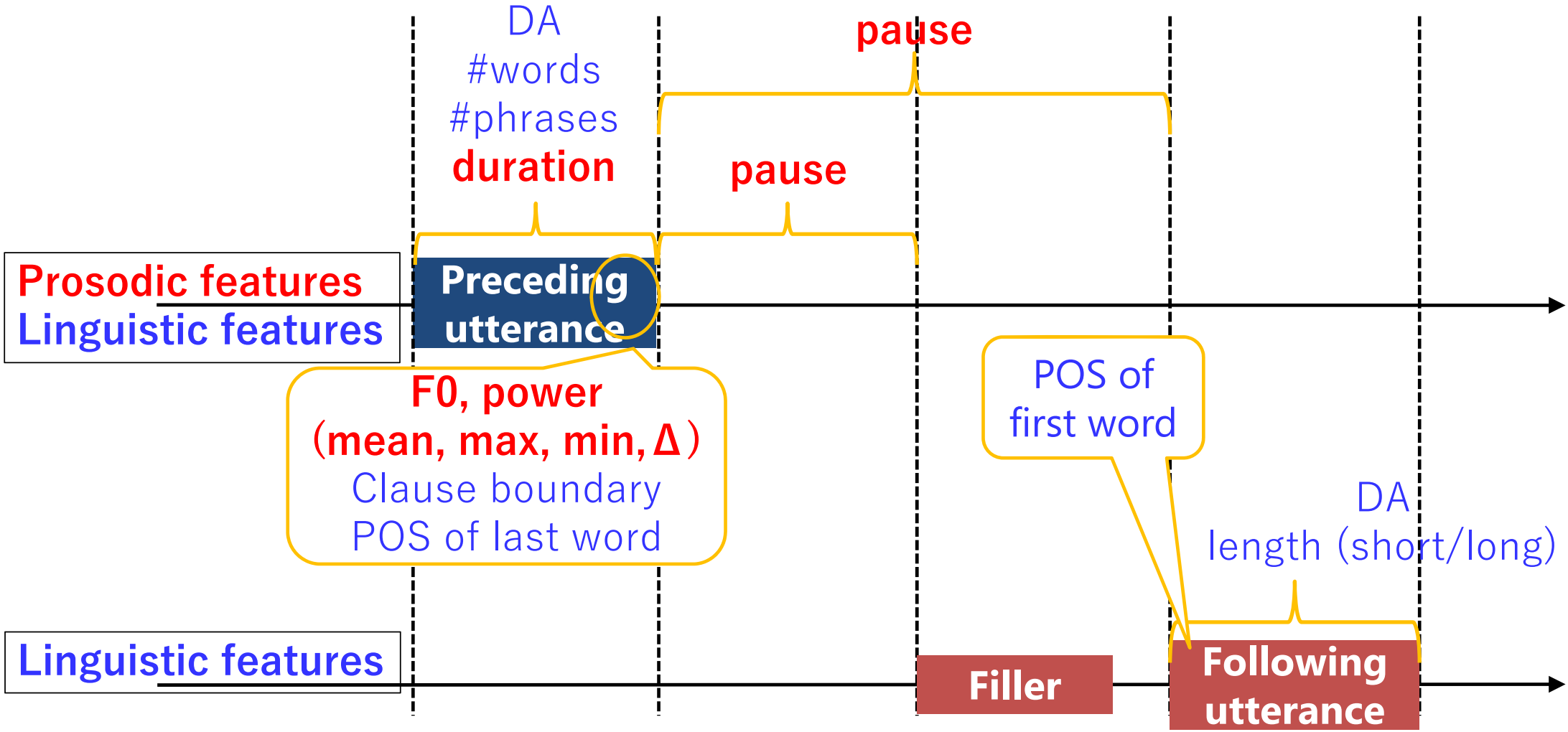
- **Demonstrative forms あの (so) are used in turn-keep**
  - Hold the turn to take time before speaking the next utterance
- **Notice forms あ (well) are used in turn-switch**
  - Indicate acknowledgment to the preceding utterance
- **Proper forms えっと (um) are used in moving to next dialog segment**
  - Hold the turn while thinking about the next utterance

# Prediction of Fillers

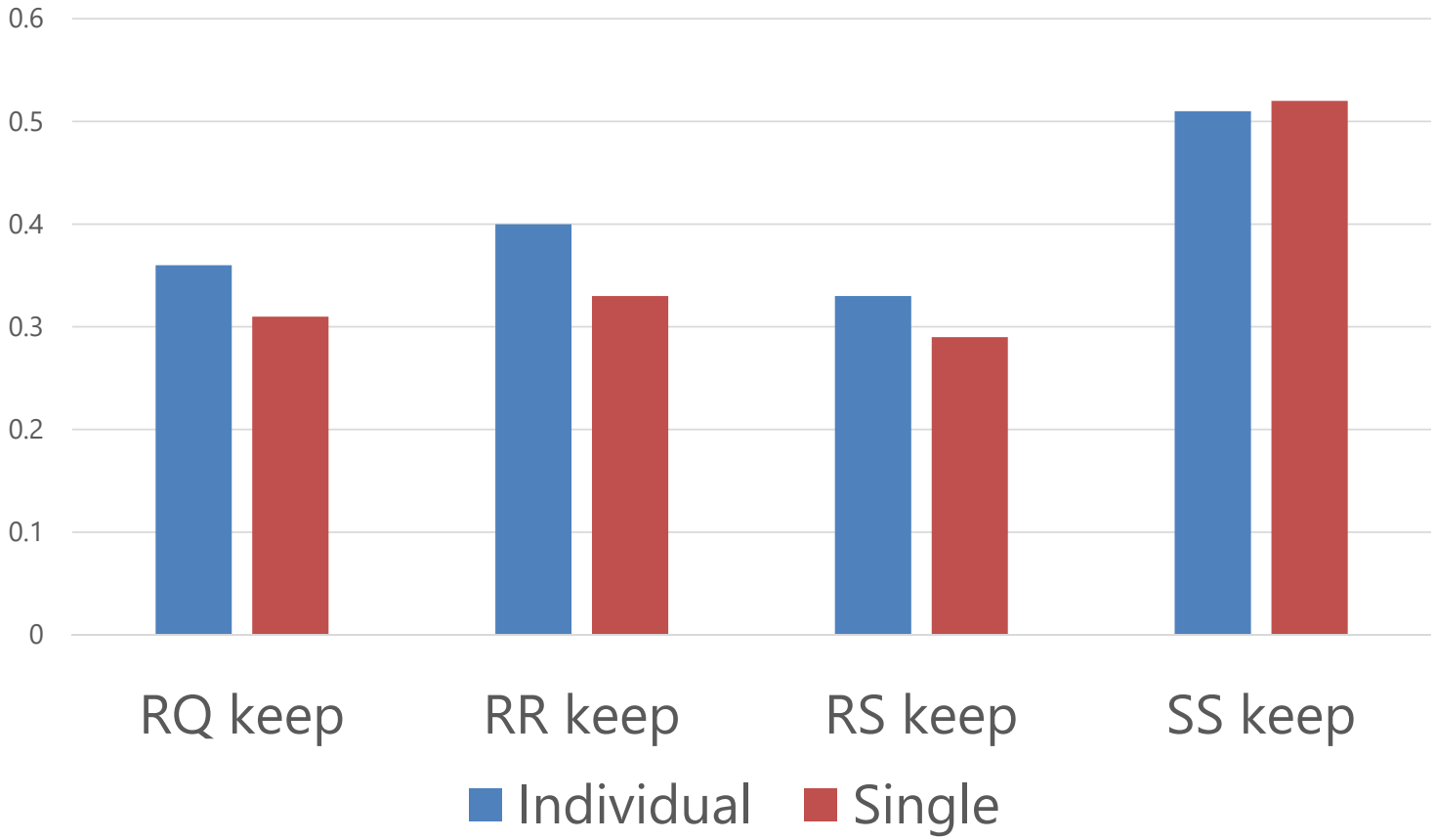
- **Given preceding DA and Following DA (DA pairs)**
  - Fillers are generated after determining to speak
- **Class of fillers**
  - Typical filler forms according to the DA pair
  - Other filler forms
  - No fillers
- **Classifier**
  - Random Forest
- **Baseline**
  - Single model to predict only occurrence of fillers by using all data, and output the typical filler

- **5-fold Cross Validation**
- **F-measure: harmonic mean of Recall & Precision**

# Features

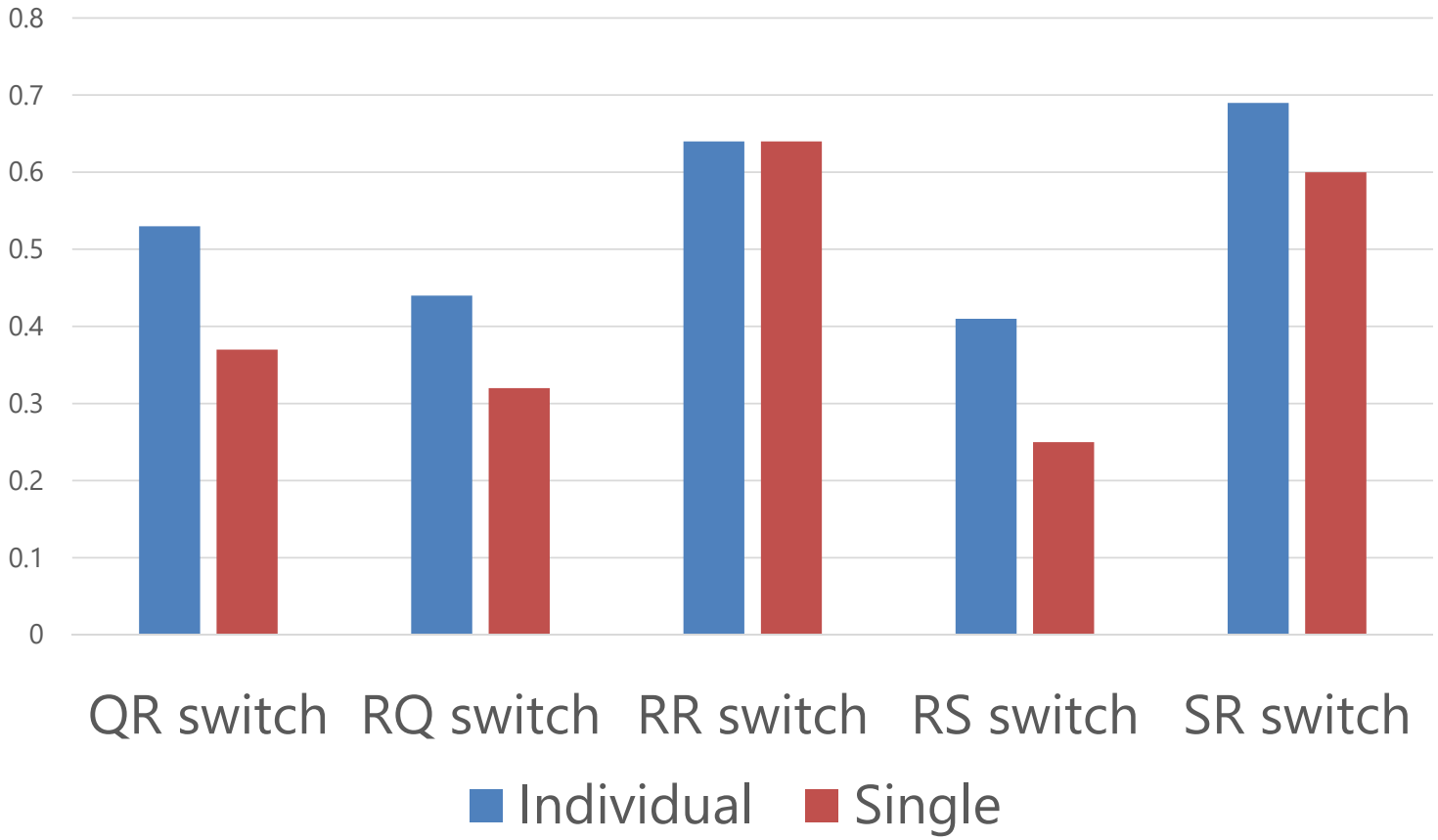

DA
#words
#phrases
duration
pause
pause
Prosodic features
Linguistic features
Preceding utterance
F0, power
(mean, max, min, Δ)
Clause boundary
POS of last word
POS of first word
DA
length (short/long)
Linguistic features
Filler
Following utterance

# Prediction Accuracy (Turn-keep; F-measure)




- **Individual modeling is more effective than single model**
- **All features (prosodic/linguistic, preceding/following) are useful**

# Prediction Accuracy (Turn-switch; F-measure)




- **Individual modeling is more effective than single model**
- **Linguistic features of following utterances are useful**

# Summary of Prediction Result

- **Prosodic features of preceding utterances are effective in predicting fillers in turn-keep.**
  - Speaker suggests turn-holding with prosody of the utterance.
- **Linguistic features of following utterances are indispensable in predicting fillers in turn-switch.**
  - Speaker generates a filler after deciding to take a turn.

- **Filler prediction model individually trained for each DA pair achieves better performance than the single model for all.**
- **Overall performance is not so high.**
  - Arbitrary characteristics of fillers, i.e. fillers may be placed or not depending on the person and at different times.

# Prediction in Speech Collision Cases

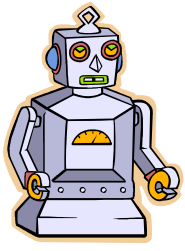
- **95 speech collisions (over 500 ms) in the corpus**
  - Ratio of filler occurrence: 25% (24/95)
- **Proposed model predict fillers in 58% (55/95)**
  - Much higher than the average (35%)
- **Potentially avoid speech collisions**
  - Collision of fillers with the user is not so harmful.
  - SDS cannot usually cancel the speech output once generated.

# Speech Collision due to Ambiguity in Turns

user

Turn-hold ⟷ Turn-switch

ambiguous

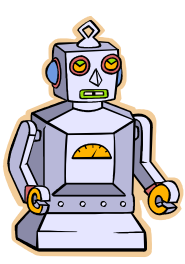


話者A ： ちょっと戻りが何時ぐらいになるかがわからないのですが、少し、あの、こちらでお待ちいただけますか。
話者B ： はい。わかりま。え、 [じゃー
話者A ： [今日は

→ ASR errors & Dialog breakdown

# Turn-holding/taking by fillers





システム　：　ちょっと戻りが何時ぐらいになるかが
　　わからないのですが、少し、あの、
　　こちらでお待ちいただけますか。
ユーザ　　：　はい。わかりま。え、［じゃー
システム　：　　［**あのー**
　　今日はどちらからいらしたんですか。

**Collision with fillers is less harmful**

# Subjective Evaluation of Generated Fillers

- **10 audio samples**
  - With many fillers
  - With some fillers
  - No filler
- **20 subjects listened and answered questionnaire**
  - Naturalness
  - Likability

- **Significant difference between some fillers and no fillers**
  - Generating some fillers improves naturalness.
  - Speech without any fillers is not natural (too artificial).
- **No difference between many fillers and no fillers**
  - It is not good to generate fillers too much.

# Conclusions

**Investigate effect of fillers for smooth turn-taking in SDS**

1. **Analysis of filler occurrence and forms in dialog act (DA) pairs**
   - Do fillers occur more often when turn-keep/switch is ambiguous? → **YES**
   - Are there specific types of fillers depending on DA pairs? → **YES**

2. **Prediction of fillers**
   - How well can we predict filler occurrence and forms? → **not so well**
   - Effect of DA pairs? → **YES**
   - Effective features? → **Prosody of preceding utterance & DA of next**

3. **Generation of fillers**
   - Does it help avoid speech collisions? → Possibly
   - Naturalness and Likability? → enhanced, but should not generate too much