英日同時通訳音声の音声認識

大村絵梨[†1]　南條浩輝[†1]

英日同時通訳音声を対象とした音声認識について述べる．これまでの話し言葉の音声認識は，主に会話や講演などの自然な音声を対象としていた．英日同時通訳では，通訳者は英語音声を聞きながらその場で日本語発話を行う．このため，同時通訳された日本語発話は通常の日本語発話と性質が異なる．このような音声の認識についての検討は不十分であるため，本論文では，英日同時通訳音声の認識を行なった．その際，多言語同時音声認識を適用した．多言語同時音声認識は，ある言語の音声とそれに対応する他言語の音声を，翻訳モデルを用いてお互いに情報を補いあい，同時に認識する枠組みである．本論文では，英日同時通訳音声を用いて多言語同時音声認識の評価も行なった．日本語音声認識時に対訳となる英語情報を利用すること，および英語音声認識時に対訳となる日本語情報を利用することの効果と問題点を示した．

Automatic Speech Recognition of English-Japanese Simultaneous Translation Speech

Eri OHMURA[†1] and Hiroaki NANJO [†1]

This paper addresses automatic speech recognition (ASR) for English-Japanese (E-J) simultaneous translation speech. Conventional ASR studies for spontaneous speech mainly target conversational speech and lecture speech, which are natural. On the other hand, in E-J simultaneous translation, the E-J translator makes a speech on the fly while listening to the English speech. For this reason, a simultaneously E-J translated Japanese utterance differs from a natural Japanese utterance. However, studies about ASR of such simultaneous translated speech are not enough. Based on the background, we study ASR of E-J simultaneous translation speech. Firstly, we show the results of conventional ASR systems, and then the results of proposed bilingual ASR system. In the bilingual ASR, corresponding E-J utterances are recognized simultaneously and complementarily with statistical ASR models and machine translation models.

1. はじめに

英日同時通訳音声を対象とした音声認識について述べる．これまでの話し言葉の音声認識は，主に会話や講演などの自然な音声を対象としていた．これに対し，本研究では英日同時通訳音声を研究対象とする．英日同時通訳では，通訳者は英語音声を聞きながらその場で日本語発話を行う．このため，同時通訳された日本語発話は通常の日本語発話と性質が異なる．このような音声の認識についての検討は不十分であるため，本論文では，このような英日同時通訳音声の認識について述べる．その際，多言語同時音声認識を適用したのでその結果についても述べる．多言語同時音声認識とは，ある言語の音声とそれに対応する他言語の音声を翻訳モデルを用いてお互いに情報を補いあって同時に認識する方法であり，理想的な条件（読み上げ音声とその対訳テキスト）において有効性が確認されている．本論文では，実際の英日同時通訳音声を用いて多言語同時音声認識の評価を行なった．

2. 英日同時通訳音声

評価データには，同時通訳音声データベース CIAIR-SIDB[1)] に含まれている 1999 年に収録された独話の英日同時通訳音声のうち政治 6 件と経済 4 件の合計 10 件を用いた．詳細を表 1 に示す．英語音声は講演音声であり，日本語音声は英語音声を同時通訳した音声である．英語，日本語ともに母語話者 10 名が発話している．

英日同時通訳された日本語音声には以下の特徴が含まれる[2)]．

1. 順送りによる訳出
   英語話者の発話の生起順に発話する
2. 短縮による訳出
   短い語への言換え（要約），重要でない箇所の省略を行って発話する

また，同時通訳音声には図 1 に示すように，通常の講演音声とは異なる言語的特徴が存在する．図 1 の A は，“ 一つの政府 ”と“ 一人の男の子 ”を通訳する順番に迷い，何度も言い直している例である．B は，余分な文の後半部分や前半部分が加わり，1 発話として見た場合，語順が通常の講演音声とは異なる例である．同時通訳音声でない講演音声で学習したモデルを用いることが適切か明らかでないが，本研究では，日本語話し言葉コーパス（corpus

†1 龍谷大学理工学研究科
Graduate School of Science and Technology, Ryukoku University

A．すすそれはあ男の一つの政府を作り上げるということは一つの男一人の男の子の夢でございましたが
B．集まって友達の家でこの放送を聞きました次の年私たちはあ作文をきまして悲惨な戦争の終わりについて書きました誰かが

図 1　通常の講演音声とは言語的特徴が異なる同時通訳音声の例

表 1　評価データ

講演 ID	英語の講演時間（分:秒）	単語数		ドメイン
		日本語	英語	
1	13:03	2,287	1,994	政治
2	7:27	926	652	
3	24:36	4,727	2,820	
4	7:08	1,291	799	
5	28:18	4,765	3,605	
6	19:56	2,564	1,944	
7	12:12	2,310	1,485	経済
8	6:53	1,267	735	
9	4:01	737	392	
10	28:55	4,781	3,477	
合計	152:29	25,655	17,555	

of spontaneous Japanese: CSJ)[3] の講演音声で学習したモデルを用いて同時通訳音声の音声認識を行う．これは他に適切なデータが見当たらないためである．

3. 音声認識システム

3.1 日本語音声認識システム

本研究では，以下に示す音響モデルと言語モデルおよび音声認識エンジン Julius rev.4.1.5（standard 版）[4] を用いて音声認識システムを構築した．音響モデルには，CSJ 付属の 2496 講演から学習した性別非依存 HMM（状態数約 3000，混合数 16 のトライフォンモデル）を用いた．また，特徴量には MFCC （Mel Frequency Cepstrum Coefficient）12 次元 + Δ 12 次元 + Δpower の計 25 次元を用いた．言語モデルには，CSJ の 2702 講演から学習した語彙サイズ 31k の逆向き単語 3-gram モデルを用いた．これらを表 2 にまとめる．

表 2　ベースライン日本語音声認識システム

音声認識エンジン	Julius rev.4.1.5（standard 版）[4]
音響モデル	性別非依存トライフォンモデル（3000states，16mixtures）（CSJ 付属の 2496 講演から学習）
言語モデル	逆向き単語 3-gram モデル（CSJ の 2702 講演から学習）
単語辞書	約 31k

表 3　ベースライン英語音声認識システム

音声認識エンジン	Julius rev.3.5.3-multipath 版（standard 版）
音響モデル	性別非依存トライフォンモデル（7000states，16mixtures）（ISTC の英語版音声認識キットに含まれる Wall Street Journal（新聞記事）から学習）
言語モデル	逆向き単語 3-gram モデル（Daily Yomiuri の英語記事（1990 年と 2000〜2005 年，総単語数 30G）から学習）
単語辞書	約 22k

3.2 英語音声認識システム

本研究では，英語音声に対しても音声認識を試みた．なお，英日同時通訳音声における英語音声は，原言語側であるので自然な講演音声と考えられる．

本研究では，以下に示す音響モデルと言語モデルおよび音響認識エンジン Julius rev.3.5.3-multipath 版（standard 版）を用いて音声認識システムを構築した．音響モデルには，ISTC の英語版音声認識キットに含まれる Wall Street Journal（新聞記事）から学習した性別非依存 HMM（状態数約 7000，混合数 16 のトライフォンモデル）を用いた．また，特徴量には MFCC（Mel Frequency Cepstrum Coefficient）12 次元 + Δ12 次元 + ΔΔ12 次元 + Δpower の計 38 次元を用いた．言語モデルには，Daily Yomiuri の英語記事（1990 年と 2000〜2005 年，総単語 30G 単語）から学習した語彙サイズ 22k の単語 3-gram モデルを用いた．これらを表 3 にまとめる．

4. 音声認識実験

10 件の同時通訳音声の日本語および英語を音声認識した．その際，音声は 0.5 秒の無音区間であらかじめ区切って認識した．結果を表 4 に示す．WER は 41.39%であった．今回用いたものとほぼ同じ音声認識システムを使用した CSJ テストセットの音声認識精度[5] と比べて，精度は低めと考えられる．ただし言語モデルの単位が異なるため正確な比較ではないこと，およびテスト音声の話題が CSJ 講演と話題が一致しない政治・経済であることが

表 4　英日同時通訳音声の認識精度（WER（%））

講演 ID	日本語 MLLR なし/あり	英語 MLLR なし/あり
1	42.46 / 37.05	64.99 / 54.49
2	36.71 / 32.69	34.10 / 28.42
3	45.29 / 34.62	28.29 / 17.88
4	47.18 / 38.72	50.44 / 32.67
5	38.43 / 31.37	29.49 / 20.29
6	32.77 / 27.52	35.00 / 19.92
7	55.13 / 49.82	43.76 / 33.99
8	47.34 / 38.41	53.90 / 33.83
9	38.75 / 30.15	52.17 / 41.43
10	36.08 / 30.08	41.26 / 21.69
平均	41.39 / 34.24	39.93 / 26.96

精度が低い原因の可能性もあることに注意する必要がある．同時通訳発話に起因する音声認識の問題については，詳細に調査を行い，明らかにする必要がある．

次に，音響モデルの教師なし適応（MLLR 適応）を行なった．具体的には音声認識を行って得た認識結果から音素ラベルを作成し MLLR 適応を行った．これを 3 回繰り返した．この話者依存音響モデルを用いて音声認識した結果も表 4 に示されている．日本語について 34.24%の WER が，英語について 26.96%の WER が得られた．同時通訳音声である日本語については精度が十分でなく，改善の余地が大きいことがわかった．

5. 英語情報を利用した日本語音声の認識

同時通訳音声である日本語の音声認識について改善の余地が大きいことがわかった．本章では，英語の音声情報を利用した日本語音声について述べる．

5.1 多言語同時音声認識の枠組み

我々はこれまでに多言語同時音声認識の枠組みを提案している[6]．英日同時通訳音声の同時認識の概要を図 2 に示す． この枠組みでの英日同時通訳における日本語の音声認識は，日本語音声 X および対応する英語音声の認識結果 E から日本語文字列 $\hat{J}$ を求める問題として定式化できる（式 (1)）．なお，N_J は日本語文字列 J の長さ（単語数）である．

$$\hat{J} = \underset{J}{\mathrm{argmax}} \left(\log P(X|J) + \alpha \log P(J) + \beta N_J + \gamma (\log P(J|E) + \delta N_J) \right) \quad (1)$$

式 (1) は，日本語の音響モデルスコア $P(X|J)$ と言語モデルスコア $P(J)$ から構成される日本語の（対数）音声認識スコア $\log P(X|J) + \alpha \log P(J) + \beta N_J$ と翻訳モデルスコア $P(J|E)$ からなることがわかる．また，英語の音声認識は，英語音声 Y および対応する日本語音声の認識結果 J から英語文字列 $\hat{E}$ を求める問題として定式化できる（式 (2)）．

$$\hat{E} = \underset{E}{\mathrm{argmax}} \left(\log P(Y|E) + a \log P(E) + bN_E + c(\log P(E|J) + dN_E) \right) \quad (2)$$

式 (1) は，英語の音響モデルスコア $P(Y|E)$ と言語モデルスコア $P(E)$ から構成される英語の（対数）音声認識スコア $\log P(Y|E) + a \log P(E) + bN_E$ と翻訳モデルスコア $P(E|J)$ からなることがわかる．また $P(E|J)$ をベイズ則にしたがって展開し，$P(J)$ が E による最大化に無関係であることを利用すると，式 (1) は，以下のとおりとなる．

$$\begin{aligned}
\hat{E} &= \underset{E}{\mathrm{argmax}} \left(\log P(Y|E) + a \log P(E) + bN_E + c(\log \frac{P(J|E)P(E)}{P(J)} + dN_E) \right) \\
&= \underset{E}{\mathrm{argmax}} \Big(\log P(Y|E) + a \log P(E) + bN_E \\
&\qquad + c(\log P(J|E) + \log P(E) - \log P(J) + dN_E) \Big) \\
&= \underset{E}{\mathrm{argmax}} \left(\log P(Y|E) + (a-c) \log P(E) + bN_E + c(\log P(J|E) + dN_E) \right) \quad (3)
\end{aligned}$$

ここで，$a - c$ を α，b を β，c を γ，d を δ とおくと，式 (4) が得られる．

$$\hat{E} = \underset{E}{\mathrm{argmax}} \left(\log P(Y|E) + \alpha \log P(E) + \beta N_E + \gamma (\log P(J|E) + \delta N_E) \right) \quad (4)$$

このことは，日本語情報を用いた英語音声認識時に英日方向の翻訳モデルをそのまま利用す

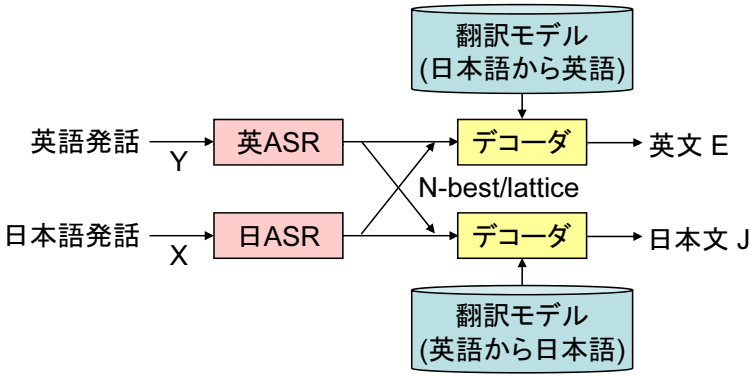


図 2　英語音声と日本語音声の同時認識の概要

表 5　対応英語発話が与えられた条件での同時通訳日本語音声の音声認識（WER（%））

講演 ID	ベースライン	英語情報あり
1	37.05	**36.74**
2	32.69	**31.94**

ることができることを示している．

5.2　多言語同時音声認識の実験条件

3 章で述べた各言語の音声認識システムと，日英新聞記事対応づけデータ[7] の 179k 文対で学習した IBM-model3 翻訳モデル（英日方向: $P(J|E)$ を与えるモデル）を用いて実験を行った．具体的には，音声認識システムにおいて N-best 候補（N=300）を出力し，それらを翻訳モデルスコアを用いてリスコアした．翻訳モデルスコア算出時には近似計算[8] を行い，日本語音声認識時のパラメータは $\alpha = 8, \beta = -4, \gamma = 1, \delta = 1$ とし，英語音声認識時のパラメータは $\alpha = 18, \beta = -12, \gamma = 1, \delta = 1$ とした．

5.3　英日発話の対応が与えられた条件での認識

多言語同時音声認識には，同一内容発話の対応が与えられている必要がある．はじめに，2 件の同時通訳音声について，人手で正解対応を与えて英語情報を用いた日本語音声認識を行った．ここでは，日本語発話に対して一つまたは複数の英語発話を正解として対応づけた．結果を表 5 に示す．2 件それぞれで WER の改善が得られた．講演 1 については，対訳英語の WER が 54.49%と高いにもかかわらず，その情報を用いる効果がみられた．

5.4　自動推定した英日発話対応を用いた認識

次に，同一内容発話の対応付けを自動で行い，それに基づいて同時認識の枠組みを適用した結果を示す．

5.4.1　自動対応付けアルゴリズム

本項では表 1 の英日同時通訳音声中の同一内容の英日発話の対応付けについて説明する．本研究での自動対応づけアルゴリズムは図 3 のとおりである．これは英日同時通訳音声では，対応する日本語よりも先に英語が話されている（図 4(a)[9]）という仮定に基づき，日本語発話に対して一つまたは複数の英語発話を対応づけるものである．

5.4.2　対応づけの結果

評価データ 10 件中，2 件（講演 1 と 2）の対応づけ評価を行なった．ここでは，日英発話を 1 対 N で対応づけるため，完全一致率（人手対応づけ結果と完全に一致する場合）と部分一致（人手対応づけの一部が自動対応づけ結果に含まれている場合）で評価した．結果

（1）　0.5 秒以上の無音で英語と日本語の音声を区切る．先頭から順に日本語発話を $\mathrm{ja}_1, \mathrm{ja}_2, ...$，英語発話を $\mathrm{en}_1, \mathrm{en}_2, ...$ とする．
（2）　$n = 1$ とする．
（3）　ja_n の発話開始時刻 T_{ja_n} よりも過去に開始された英語発話のうち，最も直近の英語発話 en_m を ja_n に対応づける．ただし，ja_n の発話長が 0.95 秒以下のときは en_m を ja_{n+1} に対応づける．ja_{n+1} の発話長が 0.95 秒以下のときは対応づけない．
（4）　n を一つ増やす．ja_n がなければ手順 5 へ，あれば手順 3 に戻る．
（5）　英語発話 $\mathrm{en}_m (m = 1, 2, ...)$ について，対応先の日本語発話がない en_m については，その発話開始時刻 T_{en_m} に最も発話開始時刻が近い日本語発話と対応づける．

図 3　自動対応づけアルゴリズム

表 6　英日発話の対応付けの自動推定精度

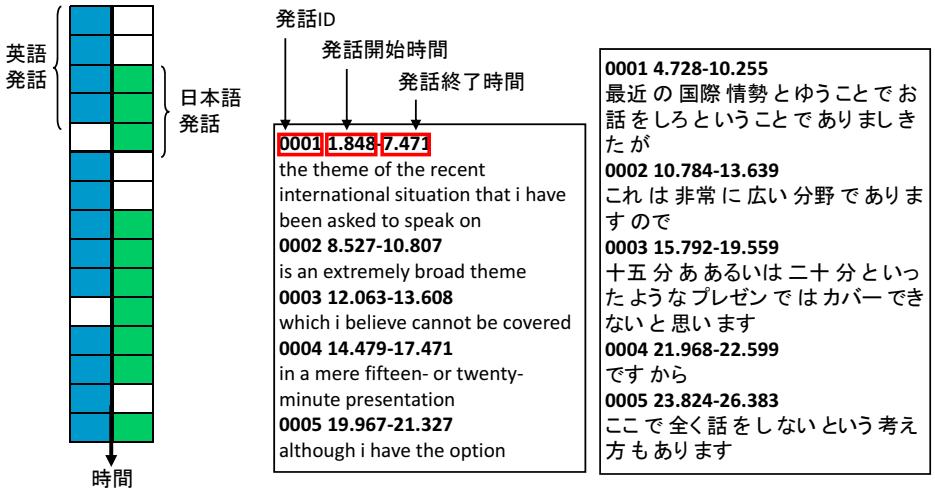
講演 ID	完全一致（%）	部分一致（%）
1	64.4	100
2	37.0	100

を表 6 に示す．講演 1，2 それぞれでの完全一致率は 64.4%，37.0%であり，部分一致率はいずれも 100% であった．すなわち，全ての日本語発話に対応する英語発話が一部または全部含まれていることを示している．

5.4.3　音声認識結果

日本語の音声認識時に英語情報を用いたときの結果を表 7 に示す．英語の音声認識時に日本語情報を用いたときの結果を表 8 に示す．

表 7 より，自動推定した対応付けと英語の音声認識結果を用いた日本語音声認識が効果的であることがわかる（10 件中 5 件で精度向上，1 件で変化なし）．さらに，表 8 より，自動推定した対応付けと日本語の音声認識結果を用いた英語音声認識が効果的であることがわかる（10 件中 6 件で精度向上，2 件で変化なし）．精度低下がみられた講演もあるが，この原因として，英日発話の自動対応づけの精度が十分でなかった可能性が考えられる．同時通訳者によっては，事前原稿に基づいて発話が開始される前に同時通訳を発話することもあ



(a) 英日同時通訳の発話の流れ　(b) 英日同時通訳の評価データ(例)

図 4　英日同時通訳音声の流れとデータ例

表 7　日本語音声認識の比較結果（WER(%)）

	英語情報なし	英語情報あり
1	37.05	**36.86**
2	32.69	**32.04**
3	**34.62**	35.11
4	38.72	**38.50**
5	31.37	**31.10**
6	**27.52**	28.05
7	49.82	**49.65**
8	38.41	38.41
9	**30.15**	30.50
10	**30.08**	30.52
平均	34.24	34.36

表 8　英語音声認識の比較結果（WER(%)）

	日本語情報なし	日本語情報あり
1	54.61	**54.06**
2	28.83	**28.53**
3	**18.94**	19.04
4	32.67	**31.91**
5	20.75	**20.69**
6	20.11	**19.80**
7	34.28	34.28
8	34.01	34.01
9	41.84	**40.31**
10	**22.63**	22.84
平均	27.49	27.36

り，このような場合には，今回用いた英日発話開始時刻の前後関係の情報のみに基づく自動対応付けは機能しない．

次に，日本語の音声認識時に英語情報用いたこと（多言語同時認識）により得られた実際の認識結果の改善例を図 5 と図 6 に示す． 図 5 の例では，対訳（英語）の認識結果に

正　解　文: 一つ目にグローバル経済それから二つ目に雇用問題三つ目に
対訳正解文: international crime and third the global economy
対訳認識文: international crime instead global economy

英語情報なし: 一つ目にグローバル**掲載**するか二つ目に**この**問題三つ目に
英語情報あり: 一つ目にグローバル**経済**するか二つ目に**雇用**問題三つ目に

図 5　英語情報を用いた日本語音声認識の改善例

正　解　文: え医薬品それから食料の不足といったところに
対訳正解文: spurred by shortages of food and medical supplies
対訳認識文: spurred by shortages of food and medical supplies

英語情報なし: えっ医薬品それから食料の**過不足**といったところに
英語情報あり: えっ医薬品それから食料の**不足**といったところに

図 6　英語情報を用いた日本語音声認識の改善例

“economy” が含まれているため，英語情報を用いることで “掲載” が “経済” と修正され，精度が向上した．図 6 の例では，対訳認識文に “shortages” が含まれているため，英語情報なしでの音声認識時には “過不足” とされた箇所が，英語情報を用いることで “不足” と修正された．

5.4.4　N-best 候補での日本語音声認識精度向上期待値

今回は，翻訳モデルを用いて N-best リスコアリングを行なうことで，多言語同時音声認識を行なった．ここで用いた日本語音声認識結果の N-best 候補文中に，精度が向上する認識結果が含まれていたかを調査した．具体的には，複数の認識結果候補文（例えば，10-best 候補）中から最も正解文に近い文を認識結果として選択できたときの精度（N-best リスコアリングによる精度向上の上限値）を調査した．結果を表 9 に示す． 10 件すべての評価データにおいて 1-best 候補より 10-best 候補の中から正解文に最も近い認識結果候補文を選択できることがわかる．すなわち，生成されている N-best 候補にはより単語誤り率が低くなる文が含まれており，正しい日英対応があれば翻訳モデルにより改善できる可能性があ

6

表 9 日本語発話の N-best 認識率（WER(%)）

	1-best	10-best
1	37.05	34.26
2	32.69	25.81
3	34.62	35.23
4	38.72	43.45
5	31.37	30.85
6	27.52	26.40
7	49.82	48.14
8	38.41	37.96
9	30.15	33.11
10	30.08	28.72
平均	34.24	33.55

ること，すなわち同時音声認識を用いてより高精度な日本語音声認識を行うことができる可能性があることがわかった．

6. おわりに

英日同時通訳が行われた講演音声 10 件を対象として音声認識を行った．同時通訳日本語発話に対して，従来の音声認識システムで 34.2%の WER が得られた．英語発話に対して，従来の音声認識システムで 27.5%の WER が得られた．日本語音声認識時に対訳となる英語情報を利用すること，および英語音声認識時に対訳となる日本語情報を利用すること（多言語同時音声認識）の有効性を調査した．認識精度の向上が得られる講演の数が，精度低下する講演の数（提案手法による悪影響）よりも多いことがわかった．対訳の自動対応付けの課題が大きいことを確認した．

謝　　辞

本研究は科研費若手研究（B）（21700210）の助成を受けたものである．

参 考 文 献

1) 松原茂樹，相澤靖之，河口信夫，外山勝彦，稲垣康善：同時通訳コーパスの設計と構築，通訳研究， No.1, pp.85–102 (2001).
2) 遠山仁美，松原茂樹：同時通訳コーパスを用いた通訳者の訳出パターンの分析，電子情報通信学会技術報告 TL2003-26，pp.13–18 (2003).
3) 小磯花絵，前川喜久雄：『日本語話し言葉コーパス』の設計の概要と書き起こし基準について，情報処理学会研究報告，NL-143, pp.41–48 (2001).
4) 河原達也，　李晃伸：連続音声認識ソフトウエア Julius，人工知能学会誌，Vol.20, pp. 41–49 (2005).
5) Kawahara, T., Nanjo, H., Shinozaki, T. and Furui, S.: Benchmark Test for Speech Recognition using the Corpus of Spontaneous Japanese, *Proc. SSPR*, pp.135–138 (2003).
6) 南條浩輝：多言語音声の同時認識枠組みの提案，情報処理学会論文誌， Vol.49, No.12, pp.4044–4048 (2008).
7) Utiyama, M. and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp.72–79 (2003).
8) Ohmura, E. and Nanjo, H.: A Study of Translation Models and Score Calculation on Bilingual ASR Framework, *Proc. APSIPA ASC 2010*, pp.530–533 (2010).
9) Toyama, H., Matsubara, S., Ryu, K., Kawaguchi, N. and Inagaki, Y.: CIAIR Simultaneous Interpretation Corpus, *Proceedings of the oriental chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques for Speech Input/Output (Oriental COCOSDA)* (2004).