放送音声の書き起こしに関する検討

西村雅史　伊東伸泰
日本アイ・ビー・エム 東京基礎研究所
e-mail: nisimura@trl.ibm.co.jp

1. はじめに

近年，ニュース音声を対象とした音声認識，特にディクテーションの研究が盛んである．この技術が確立されれば，字幕の付与，情報検索，話題抽出，Audio Indexingなどへ幅広い応用が期待できる．実際，米国では実用化を前提とした試行が始まっているし，日本でも字幕の自動化を目的として活発に研究が行われている．

ここでは，ニュース音声書き起こしシステムを構築する上での課題について考察するとともに，試作したシステムについて紹介する．

2. 実際のTVニュース音声の調査

2-1. 調査結果

表1および表2に，東京放送(TBS)の30分間のTVニュース(ニュース1130，'97年10月30日にオンエアされたもの)を調査した結果を示す．「読み上げ」はスタジオ収録されたアナウンサー(アンカーマン以外の話者を含む)による原稿の読み上げを指す．記者レポートにも原稿がある場合が多いが，少なくとも全173文の発声中71文は原稿を読み上げたものではなかったと思われる．これには街頭インタビューや，記者会見，天気概況などが含まれる．

実に全体の56%の文章(96文)に意図的に背景雑音(効果音など)が重畳されている．音声品質が劣化するスタジオ外からの記者レポートもさる事ながら，このような効果音は人間の声，音楽など，定常雑音ではないケースがほとんどであり，その上信号レベルも高く，認識を大変困難なものにしている．実際この96文のうち，人間の声が効果音として重畳されていたケースが36文，音楽が重畳されたケースも33文あった．

間投詞（ここでは文字としては表現しにくい「うなり声」のようなものも間投詞に含む）は全部で119回観測され，全体の45%にあたる77文には何らかの間投詞が含まれていた．このデータを見る限り未知語よりも頻度が高く，適切な対策が必要であることが分かる．なお，観察された全不要語のうち80%(=95回)は"エ"および"エー"という発話であった．一方，未知語については，60K語彙の辞書を用いた場合の未知語率は1.4%であり，タスクによらず，新聞などと大差ないレベルとなっている．なお，未知語のうち約80%(42単語)は人名であり，基本語彙に網羅的に追加することは現実的ではない．

表1 TVニュースの調査結果：効果音の重畳

タスク	*発声文数*	*効果音が重畳された文*
読み上げ	65	22 (34%)
記者レポート	37	25 (68%)
その他の発話	71	49 (69%)
全体	173	96 (56%)

表2 TVニュースの調査結果：間投詞および未知語

タスク	*単語数*	*間投詞*	*未知語*	*Perplexity*
読み上げ	1953	45	32 (1.6%)	151.5
レポート	992	22	6 (0.6%)	208.1
その他	842	52	15 (1.8%)	440.1
全体	3787	119	53 (1.4%)	208.7

間投詞は入力単語に含まず．語彙サイズ60K．
Perplexityは未知語以外の部分から推定

言語的な側面からは，たとえ原稿を読み上げている場面でも，既存の新聞記事データベースなどとは表現，特に言い回しが違うことが問題になると予想された．しかし，今回，後述する汎用60Kの言語モデルを適用したところでは，表2に示すように「読み上げ文」や「記者レポート」に対する単語Perplexityは新聞タスクと比べて大差ないものであることが分かった．これは言語モデルの学習時に，パソコン通信等から得られた口語体の文章を多数用いていた効果だと考えている．一方，自由発話に近い「その他の発話」の場合には，Perplexityは440にも達しており，この言語モデルで対応できているとは言い難い．

2-2. 認識実験結果

実放送をVHSテープで録音した音声と，既存の汎用ディクテーションシステム[1]を用いて不特定話者認識実験を行った．予想されたとおり，「その他の発話」部分の精度が特に低く，ほとんど認識できない．また，背景効果音が付与された場面，間投詞挿入個所の前後，未知語の前後，アンカーマンが原稿に目をやって下を向きマイクと口との位置関係が変化する場面などで認識率が大きく低下した．なお，読み上げと記者レポートでは顕著な精度差は見られなかった．

3. 汎用システムに基づくニュース音声書き起こしシステムの試作

このように，特に音響上の問題から実際のTVニュースを視聴者側で書き起こすことは現状では相当難しいが，放送局側でならば効果音やマイクの問題などへの対処が可能であり，問題をずっと簡単にすることが出来る．このため，ここでは音響上の認識率劣化要因がおおむね排除されると仮定し，残された課題についての対策を検討する．

3-1. システムの基本構成

ここで用いたディクテーションシステムの基本構成は文献[1]に示したものと変わらないが，認識対象語彙を，39,295語(40K語彙)から59,782語(60K語彙)に増やすとともに，使用帯域を5kHzから8kHzに拡大し認識精度の向上を図っている．なお，このシステムはMMX Pentium-200MHz程度のCPUで実時間動作する．

3-2. ニュース用言語モデル

パソコン通信上の発言等を言語モデルの学習に使用しているため，上記システムでもある程度口語的な表現に対応できるが，ニュース特有の言い回しや時事データに対処するため，ニュース原稿データベースも併用する．ただ，現在のところニュース原稿のデータ量は新聞記事などに比べて少なく，信頼性の高い統計量を得るには不十分であるので，言語モデル推定時の語彙は先の60K語と同じとし，汎用のトライグラム言語モデルの確率と補間して使用する．

3-3. 間投詞の処理

TVニュース音声を調査した結果から，「え」，「えー」といった間投詞が多く発せられることが分かった．これらの発声については多くの場合前後に若干のポーズが入って

いたので，この情報を利用すれば，連続発声中に現れる助詞の「へ」等と区別して検出できる．具体的には間投詞を，「無音」と同様にそれ自体の出現確率は推定するが，後続の単語の予測時には存在しないものとして処理する「透過単語」として扱う．なお，透過単語の出現確率は予備実験から実験的に推定している．

3-4. 句読点の自動挿入

認識結果の可読性を高めるため，句読点の自動挿入を行う．このような機能は言語的な後処理として付与することも可能であるが，ここでは「無音」と句読点位置が対応付くことを積極的に利用する．具体的には「無音」が検出された時これを透過単語として処理するか，句読点として出力するかを言語的な尤度に基づいて決定する．

3-5. 未知語の処理

ここでは，時期差の小さい原稿等から未知語を検出[2]し，それを認識対象単語として発音辞書にだけ追加登録する方法を検討した．言語モデルではこれらの単語を未知語のクラスに属するものとして処理する．この場合，追加された単語は既存の言語モデルとは一切無関係ないので64Kという制約は受けず，発音辞書からの削除も容易である．

4. 評価用ニュース音声による認識実験

4-1.評価実験用データ

男女各3名のアナウンサーが読み上げたクリアなニュース音声（各話者共通の119文）を使って性能評価を行った．収録はすべてTBSのラジオスタジオで，放送用の機器を用いて行っており，先に示したような音響上の問題を含まない例と考えることが出来る．ただし，不特定話者用音響モデルの推定に用いた音声データとは採取環境が異なる．また，読み上げた原稿には事件や経済の報道など，新聞記事にも多く見られるタスクだけでなく，既存のコーパスだけでは対処が難しいと思われる，スポーツの対戦結果，天気予報，交通情報などを多数含む．

4-2. ニュース原稿データベース

TBSのニュース原稿（'95年から約3年分）約20万文(5.9M単語)をニュース用言語モデルの学習データとした．なお，この原稿にもスポーツ，天気予報，交通情報などの記事はほとんど含まれていない．また，汎用60K言語モデルを用いた場合，この原稿から無作為抽出した455文に対する単語Perplexityは129.0，未知語率は1.5%であった．

4-3. ニュース用言語モデルの効果

表3，4に先のTVニュースおよび評価実験用ニュースに対するニュース用言語モデルの効果を示す．TVニュースに対するPerplexityと比較すると，この認識テストセットが言語的には大変困難なタスクであったことが分かる．評価実験用ニュースに対しては，ニュース用の言語モデルだけでは汎用60Kの能力にも満たないが，3-2.で述べたように汎用の言語モデルと組み合わせれば，わずか20万文程度のデータで推定したモデルでも，Perplexity削減のためにはかなり有効に働くことが分かった．なお，この言語モデルによりTVニュース全体のPerplexityは89.6まで削減されたが，「その他」に分類されたものだけに限るとPerplexityは333.7であり，十分な効果があったとは言い難い．

4-4. 認識性能評価実験結果

先の評価実験用ニュース音声に対し，汎用の認識システムを使用した場合の認識性能を，使用した言語モデル，話者適応の有無，および，未知語の追加登録の有無に分けて表5に示す．

話者適応は各話者100文の発声を用いて行った．用いた適応化手法はMLLRおよびMAPである．一方，未知語の発音辞書への登録については，本来，認識時と時期差の小さい原稿から自動的に抽出を行うことを想定しているため未知語率が0%となる保証はないが，この実験では評価用ニュース原稿に含まれていた未知語をすべて追加登録している．

汎用の60Kモデルを用いた場合，文字誤り率(CER)は12.1%に達するが，未知文字率が5.6%であったので，誤認識の半分近くは未知語に起因していたと言える．実際，未知語を発音辞書に登録するだけで文字誤り率は40%近く削減される．さらに，話者が特定できる場合には話者適応化が有効である．一方，Perplexityにおいて大幅な改善が見られたニュース用言語モデルの併用に関しては，文字誤り率において5%程度の比較的小さな改善にとどまっている．

すべての手法を組み合わせた実験4では，実験1に比べて文字誤り率で63%，単語誤り率(WER)でも60%の削減となった．未知語がすべて発音辞書に登録され，話者も特定できるという理想的なケースではあるが，タスクは実際のニュースと同程度，あるいはそれよりも困難なものであったことを考えると音響的には予想以上に高い性能が得られているといえる．これはアナウンサーの発声の品質に負うところが大きい．

なお，間投詞については今回用いた評価実験用ニュース音声ではほとんど観察されず，提案手法の有効性を検証することは出来なかった．

表3 実際のTVニュースによるLMの性能比較

言語モデル	*未知語率*	*Perplexity*
汎用40Kモデル[1]	2.7%	275.8
汎用60Kモデル	1.4%	208.7
News用60Kモデル	1.4%	134.9
汎用60K+News用	1.4%	89.6

表4 評価実験用ニュースによるLMの性能比較

言語モデル	*未知語率*	*Perplexity*
汎用40Kモデル[1]	5.2%	456.6
汎用60Kモデル	3.8%	309.0
News用モデル	3.8%	391.1
汎用60K+News用	3.8%	221.3

表5 評価実験用ニュース音声の認識結果

実験	*言語モデル*		*話者適応*	*未知語登録*	*CER (%)*	*WER (%)*
	汎用	*News*				
1	○	×	×	×	12.1	15.5
2	○	×	×	○	7.3	9.6
3	○	×	○	○	4.7	6.5
4	○	○	○	○	4.5	6.1

5. おわりに

汎用のディクテーションシステムを利用して実時間動作可能なニュース音声書き起こしシステムを試作した．効果音や雑音の重畳，自由発話といった問題点を除けば，おおむね実用的なレベルの精度が出ているといえる．今後は自由発話により近い形態の音声が書き起こせるよう研究を進めたい．

謝辞 評価実験用音声データの採取にご協力いただくとともに，原稿データの使用を許可してくださった(株)東京放送に感謝します．

参考文献

[1] 西村他，情処研究会-SLP 20-3 (1998.2)
[2] 伊東他，情処研究会-NLP 122-9 (1997.11)