音声言語情報処理 42−11
(2002. 7. 13)

SLP雑音下音声認識評価ワーキンググループ活動報告

中村 哲[1], 武田一哉[2], 黒岩眞吾[3], 山田武志[4],
北岡教英[5], 山本一公[6], 西浦敬信[7], 藤本雅清[8], 水町光徳[1]

[1] ATR音声言語コミュニケーション研究所, [2] 名古屋大学,
[3] 徳島大学, [4] 筑波大学, [5] 豊橋技科大学, [6] 信州大学, [7] 和歌山大学, [8] 龍谷大学

あらまし 本稿では, 2001年10月に音声言語情報処理研究会内に設立した雑音下音声認識の評価に関するワーキンググループの活動状況の報告を行う. このワーキンググループでは, 雑音下音声認識に於ける評価法, 共通のコーパスの策定に加えて, 欧州で進められているETSI AURORA雑音下音声認識アルゴリズム開発プロジェクトに合わせたアルゴリズム開発を目指している.

キーワード 雑音下音声認識,AURORA

Progress Report of SLP Working Group for Noisy Speech Recognition

Satoshi Nakamura[1], Kazuya Takeda[2], Shingo Kuroiwa[3], Takeshi Yamada[4],
Norihide Kitaoka[5], Kazumasa Yamamoto[6], Takanobu Nishiura[7],
Masakiyo Fujimoto[8], Mitsunori Mizumachi[1]

[1] ATR Spoken Language Translation Research Labs., [2] Nagoya University,
[3]Tokushima University, [4]Tsukuba University, [5]Toyohashi University of Technology, [6]Shinshu University, [7]Wakayama University, [8]Ryukoku University

Abstract This paper reports current status of the SLP working group establised in October 2001 on the noisy speech recognition. The working group aims to develop standards, common corpus, and noisy speech recognition system in conjunction with Europian ETSI AURORA evaluation projects.

key words Noisy speech recognition, AURORA

1 はじめに

音声認識においては, 隠れマルコフモデルと確率言語モデル, および大量の学習データの収集により, 学習データと同一の性質のテストデータに対しては, 非常に高い認識性能が得られるようになった. しかしながら, 実際に認識装置を利用する状況での音声は, 種々のバリエーションが生じて, 学習データとは異なったものが生じてくる. このバリエーションとしては, 大きなものとして加法性, 乗法性雑音の混入, 発話スタイルの変化がある. しかし, 学習データとして集められるデータの量は, 限られているためこのような場合, 十分な性能を獲得することは難しい. 特に, 本稿では雑音の混入による音声の認識性能の劣化について述べる.

本稿では, 2001年10月に音声言語情報処理研究会内に設立した雑音下音声認識の評価に関するワーキンググ

ループの活動状況の報告を行う．このワーキンググループは，特にEUROSPEECH2001における欧州のAURORAセッションに刺激を受けて発足したものである．この欧州のAURORAプロジェクトというのは，携帯電話などとネットワークを利用した場合の音声認識サービスを実現する際に，音声認識の前処理部を標準化しようという試みであり，標準化においては，十分利用環境において高い性能を期待できる音声認識前処理が必要ということである．このネットワークで利用する音声認識は分散型音声認識(DSR:Distributed Speech Recognition)と呼ばれている．この標準化は実際には，ETSIのDSR標準化に加わっている主として企業が具体的な要求基準，技術開発，評価を進めているが，それと平行して評価データをELRA(European Language Resources Association)から一般の研究者に配布し，主にISCAのEUROSPEECH, ICSLPにおいて評価結果を発表するAURORAスペシャルセッションが開催されている．本ワーキンググループでは，このAURORAスペシャルセッションに関連して，日本において，並列に同様の雑音下音声認識の評価のためのデータ収集，評価法の検討，性能評価のための仕組みの検討，AURORAスペシャルセッションへのコミットメント，日本からの参加者のプロモーションなどを活動の趣旨としている．

以下，第2章でAURORAスペシャルセッションの評価の方法，内容などを述べ，第3章でワーキンググループの活動経過とその内容を述べる．

2 AURORAスペシャルセッション

この欧州AURORAプロジェクト[1]は，ETSIのDSR(Distributed Speech Recognition)の標準化活動[2]に同期して進められているもので，雑音に強い音声認識の前処理を開発し，標準化することを目的としている．米国で行われているDARPAのSPINEプロジェクトなどに比べて，雑音処理のみに焦点を当てるため比較的小さなTI-DIGITSの数字認識をタスクとしている[3].

2.1 AURORA2タスク

2001年のEUROSPEECHで扱われたものが英語のTIMIT連続数字データ(LDCから公開済み)を対象としたタスクである．

2.1.1 配布物

- Training Set
  - Clean-condition training: 8440発話のClean(雑音重畳のない) TIMITデータ
  - Multi-condition training: 上記データに種々の雑音を種々のSNRで混入させたもの．
    - 雑音：subway, babble, car, exhibition
    - SNR: 5,10,15,20, Clean
- Test Set
  - Set A 学習データと同一種の雑音 (subway, babble, car, exhibition)
  - Set B 学習データと異なる雑音 (restaurant, street, airport, station)
  - Set C 音響特性が異なるもの (subway, streetにMIRS特性を付与したもの)
  - Test set SNR: -5,0,5,10,15,20,Clean
- HTKスクリプト
  - 分析条件
    - Sampling frequency 8kHz
    - フレーム周期 10msec
    - フレーム長 25msec
    - Pre-emphasis 0.97
    - Feature: 12MFCC+pow+Δ+(ΔΔ)
  - HMM
    - Unit: 数字Whole Word Model 16-states, 3-mixtures
    - 無音モデル(sil): 3-states, 6-mixtures
    - spモデル: silの第2状態とtied
  - 評価モード
    - Clean-condition test: Clean学習データで学習し，テストセットA-Cで評価
    - Multi-condition test: Multi-condition学習データで学習し，テストセットA-Cで評価
    - 評価は共通のExcel spread sheetで平均の性能，ベースラインからの改善率で評価.

図1に上記分析条件に於けるベースライン結果をあげておく．参加者はこのベースライン結果に対し前処理や適応化処理を行い性能改善率を競うことになる．全体の性能改善率は，改善率の平均として与えられる．

Multicondition training, multicondition testing

	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.59	98.67	98.57	98.83	98.67	98.59	98.67	98.57	98.83	98.67	98.62	98.67	98.65	98.66
20dB	97.82	97.94	98.24	97.47	97.87	97.73	97.61	97.61	97.66	97.65	97.67	97.58	97.63	97.75
15dB	96.65	97.43	97.70	96.88	97.17	96.16	96.67	96.60	96.27	96.43	96.50	96.31	96.41	96.79
10dB	94.38	95.47	96.18	94.11	95.04	92.94	94.86	93.71	93.92	93.86	93.83	93.92	93.88	94.45
5dB	89.01	88.21	87.53	87.60	88.09	85.05	86.58	87.53	85.16	86.08	83.11	84.16	83.64	86.43
0dB	67.85	63.18	54.10	63.71	62.21	60.88	63.06	66.27	58.07	62.07	46.21	56.35	51.28	59.67
-5dB	26.56	27.33	20.22	23.63	24.44	27.11	27.66	29.91	21.75	26.61	19.22	24.73	21.98	25.71
Average	89.14	88.45	86.75	87.95	88.07	86.55	87.76	88.34	86.22	87.22	83.46	85.66	84.56	87.03

Clean training, multicondition testing

	A					B					C			
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average	Subway M	Street M	Average	Average
Clean	98.89	99.03	99.05	99.26	99.06	98.89	99.03	99.05	99.26	99.06	99.17	99.09	99.13	99.07
20dB	96.75	90.54	97.08	96.20	95.14	90.14	95.86	89.95	94.79	92.69	93.37	95.13	94.25	93.98
15dB	91.53	72.19	88.55	90.03	85.58	74.52	88.15	73.84	81.24	79.44	86.03	89.09	87.56	83.52
10dB	75.53	47.61	63.53	72.29	64.74	51.89	66.05	49.27	55.20	55.60	71.94	75.03	73.49	62.83
5dB	47.34	22.91	30.75	39.08	35.02	26.80	36.28	24.60	24.96	28.16	50.63	50.57	50.60	35.39
0dB	22.44	5.53	10.71	14.25	13.23	7.12	17.35	10.50	9.50	11.12	24.53	23.64	24.09	14.56
-5dB	10.65	0.12	6.83	6.85	6.11	0.95	8.62	5.28	6.14	5.25	12.90	11.19	12.05	6.95
Average	66.72	47.76	58.12	62.37	58.74	50.09	60.74	49.63	53.14	53.40	65.30	66.69	66.00	58.06

図 1: Baseline results for clean- and multi-condition training

Eurospeech2001 における AURORA2 評価について，主催の D.Pearce 氏が行った発表のサマリーのグラフを図 2 示す．

AURORA は，次に示すように，さらに自動車内で実収録した数字，コマンドデータの認識評価，さらには，雑音下における大語彙連続音声認識評価へと継続，発展していく予定である [4]．

2.2 AURORA3 タスク

雑音を加算するのではなく実際に雑音下で収録した音声を用いて評価を行うことを目的としている．SpeechDat-Car プロジェクトで収録した自動車内発話音声を評価に用いる．対象言語もフィンランド語，イタリア語，スペイン語，ドイツ語，デンマーク語を対象にして，言語による差を確認することができる．また，実環境で遭遇する可能性のある次の 3 つの状況を設定し評価項目としている．

1. Well matched training and testing: ハンズフリーマイクロホンで種々のスピードで走行した場合の音声で学習，テストを行う．
2. Moderate mismatch training and testing: 中程度のミスマッチがある条件，例えば，ハンズフリーマイクロホンを用いて低速走行中音声で学習し，高速走行中音声でテストを行う．
3. High mismatch training and testing: 非常に学習とテストが異なる条件，例えば，近接マイクロホンで収録した音声で学習し，種々の速度で走行中の音声をハンズフリーマイクロホンで収録したものでテストを行う．

2.3 AURORA4 タスク

Noisy WSJ-large vocabulary evaluation: Wall Street Journal データベースに雑音とフィルタ特性を付加した大語彙データベースを用いて評価を行う [5]．評価条件は，次の 4 つの平均性能である．

- 8kHz clean training
- 8kHz multi-condition training
- 16kHz clean training
- 16kHz multi-condition training

この評価のためのスクリプトはミシシッピ大学により準備されつつある [6]．

3 WG の活動内容

WG 設立以来，AURORA 性能評価に参加するグループの啓蒙，雑音下音声認識を評価するための評価法，データ設計，スクリプト，それらの短期，長期の計画などについて，下記に示す 4 回の会合を通して議論を進めてきた．

1. 2001 年 12 月 20 日　東京工業大学
2. 2002 年 3 月 8 日　名古屋大学
3. 2002 年 3 月 19 日　神奈川大学付近
4. 2002 年 4 月 26 日　機械振興会館

これまでの活動のまとめを以下に述べる．

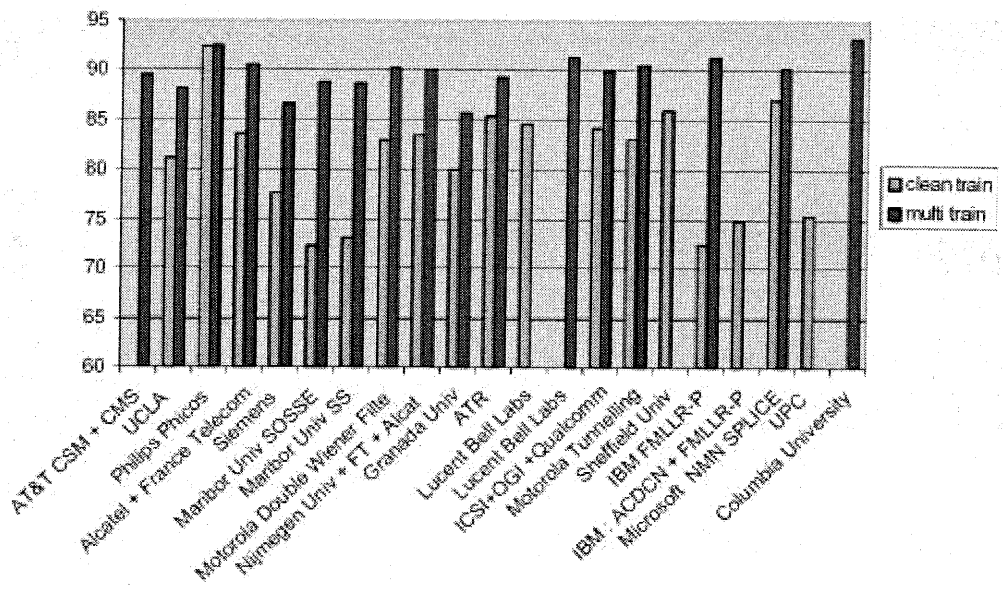

図 2: Performance comparison of Aurora2 evaluation

3.1 ICSLP2002 AURORA セッション

ICSLP の AURORA セッションについては，Organizing Committee にメンバーの一人が入り，情報の流れをよくしたほか，これまでの論文のサーベイなどもWGのメンバーで行い，調査活動も共同で行った．ICSLP2002 の AURORA セッションは，主として AURORA3 の自動車内音声データの評価を中心に行うこととなったが，WG としては AURORA2 の評価結果を投稿することとした．AURORA2 関連で約 3 件，AURORA3 に 1 件の投稿を行い，すべて採録された．

3.2 評価データの設計について

現在の AURORA タスクは，日本語が含まれていない．前処理なので，言語依存性は低いと言われているが，AURORA3 の 4 言語でも相当な認識性能の隔たりがあるので，日本語の評価データを作る必要がある．現在，次のようなデータ収録を計画している．

[AURORA2J] AURORA2 の日本語版である．TIDIGIT の数字を日本語にして，AURORA2 と同一の雑音，伝達特性を付与したものを作成する．

[AURORA2.5J] AURORA2J の単語を AURORA2 の雑音再生下で発話する．雑音の加算と実際の環境での発話の違いを考察できる．

[AURORA3J] AURORA3 と同様に自動車内発話の音声を収録する．発話内容は，孤立単語，バランス文を収録する．(名古屋大学で収録)

[AURORA4J] 大語彙の雑音下音声認識を中心としたタスクとするか，よりアプリケーションに近い環境での発話にするか，単語でなく自由発話にするかなど現在検討中である．

[雑音 DB] これまで，電子協の雑音 DB[8], Noisex92[9], RWC-DB[10] などが収録されてきたが，未だ不十分であり，種々多様な評価を行う際に基準となるような雑音データベースが必要である．

3.3 評価ツールの設計

本WGでは，AURORA と同様にデータの配布とともに評価のため，HTK のスクリプトと評価のための Spread Sheet を配布する．また，より一般性のある評価基準を確立することを目指して，種々の共通ツールの作成を目指している．ツールには，SNR 測定プログラムなどが含まれる．

3.4 アンケート

雑音DB，AURORA4Jに関連して，雑音環境として，どのような環境の雑音，残響を考慮すべきかは，非常に重要でかつ設定するのが困難な課題である．WGとしては，実際の音声認識装置の開発者，利用者，利用想定者にアンケートを行い，雑音環境の洗い出しを行いたいと考えている．

4 今後の計画とまとめ

雑音下音声認識評価に関するSLP-WGの活動の内容と現状について報告を行った．音声認識の雑音環境に於ける頑健性の問題は，今日では非常に重要な課題である．WG設立以来，AURORAの評価プロジェクトと平行して活動を進めてきた．今後，AURORAについては，ICSLPのセッションに参加し情報交換をするとともに，日本語データの収録が終了した際には，AURORAの評価DBに日本語DBを公開することなどを計画している．また，上述の日本語のデータ収集計画の議論を継続するとともに，評価手法などの検討を進めていく．WGとしては，本活動が雑音下音声認識に於ける評価の枠組みの確立の一助になればと考えている．

[謝辞] 本研究の一部は，通信・放送機構の研究委託により実施したものである．

参考文献

[1] http://eurospeech2001.org/ese/ NoiseRobust/index.html, http://www.elda.fr/proj/aurora1.html, http://www.elda.fr/proj/aurora2.html

[2] ETSI standard document, "Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm", ETSI ES 201 108 v1.1.2 (2000-04), 2000

[3] H.G.Hirsh, D.Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions", ISCA ITRW ASR2000, september, 2000

[4] D.Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & What next", Proc. EUROSPEECH2001, 2001

[5] Aurora document no. AU/337/01, "Experimental framework for the performance evaluation of speech recognition fron-ends on a large vocabulary task: Version 1.0", Ericsson, June 2001

[6] Aurora document no. AU/345/01, "Large vocabulary evaluation of front-ends- baseline recognition system description", Mississipppi State University, Aug 2001

[7] http://www2.slt.atr.co.jp/ñakamura/ SLPWG2001/ Noise-WG.html

[8] http://www.milab.is.tsukuba.ac.jp/ corpus/noise_db.html

[9] http://www.speech.cs.cmu.edu/comp.speech/ Section1/Data/noisex.html

[10] http://tosa.mri.co.jp/sounddb/index.htm