ヒューマンインタフェース 97−21
音声言語情報処理 40−21
（2002. 2. 2）

n-best 音声認識と逐次理解法によるロバストな音声理解

宮崎 昇　中野 幹生　相川 清明

日本電信電話株式会社　NTT コミュニケーション科学基礎研究所
〒 243-0198 神奈川県厚木市森の里若宮 3-1
nmiya@atom.brl.ntt.co.jp　http://www.brl.ntt.co.jp/cs/dug/

概要

本研究は対話音声の音声理解に関するものである。先に提案された逐次発話理解法 (ISSS) は、対話音声の特徴の一つである複数の発話区間にまたがって意味をなすような発話に有効であるが、音声理解の精度が低い欠点があった。本研究では ISSS に n-best 音声認識候補を適用し音声理解精度の向上を試みる。提案法は、音声認識の曖昧性と構文解析および理解状態変更規則の適用に関する曖昧性を全て理解状態の多重化により保持し、各処理における優先度を統合して各理解状態の優先度を計算しながら逐次的に理解状態を変更する。

人間-機械間の対話データに対し対話システムの理解フレーム内容を評価対象として提案法を評価し、8.49% の音声理解率向上を達成した。

キーワード: 音声理解, 対話音声,n-best, 音声対話システム

Robust speech understanding using incremental understanding with n-best recognition hypotheses

Noboru Miyazaki, Mikio Nakano, Kiyoaki Aikawa

NTT Communication Science Laboratories, NTT Corp.
3-1, Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198, Japan

Abstract

One of the linguistic features of spoken language is that the meaning can be conveyed over several utterances. The Incremental Sentence Sequence Search algorithm (ISSS) is known to be effective for this phenomenon. However, a performance of ISSS is not very good because it refers only 1-best ASR result. Therefore we use n-best recognition hypotheses with ISSS to improve speech understanding performance. Proposed method holds multiple possible belief states to represents ambiguities over syntax and semantic rule application and ASR hypotheses. Proposed method showed 8.49% improvement on speech understanding rate that is based on belief state accuracy over dialogue corpus.

keywords: speech understanding, spoken language, n-best, spoken dialogue system

1 はじめに

音声認識、言語理解および音声合成技術の発達を背景にして、音声対話システムは商用サービスが開始されるなど注目を浴びている。しかし、自由な発話を許す音声対話時の人間の発話は従来技術が対象としてきた書き言葉の読み上げ発話とは言語的な特徴が異なるため、十分な音声理解精度が得られない。

対話音声の言語的な特徴の一つとして、複数の発話区間にまたがって意味をなすような発話列が存在する。例えば「えっと会議室の予約をお願いしたいんですけど、月曜日と、」「えー、水曜日の午後 1 時」「あ、じゃなくて 3 時です」という三つの発話区間からなる入力は、月曜日と水曜日の午後 3 時に会議室を予約する、という意図を表している。ここで最初と二番目の発話区間を独立に処理すると予約日が月曜と水曜の二日間であることは自明ではなくなる。したがって、発話区間をまたがる言語処理が必要と

なる。一方、音声対話システムはユーザからの入力に応じて実時間で応答を出力する必要があるが、応答を開始すべきタイミングは一般に自明ではない。上記の例では、二番目の発話区間終了時よりも三番目の発話区間終了時の方が正しい理解結果が得られる。しかし、発話区間の間のポーズ長があまり長いと対話が不自然になるため二番目の発話終了時にシステムが主導権を取って応答を開始する方が良い場合もある。このように、応答を開始すべきタイミングは自明ではなく、各時点での入力発話内容と円滑な対話進行に関わるその他の要因に依存する。したがって、システムは各時点での応答を選択するために、発話区間毎に理解結果を決定する必要がある。これらの問題に対し、中野らは逐次発話理解法 (ISSS : Incremental Sentence Sequence Search) を用いて発話区間をまたがる言語処理と発話区間毎の理解結果の決定を両立し、その有効性を示した [4]。

一方、対話音声は言語的、音響的な特徴が読み上げ文章と異なるため音声認識の精度が低い。そこで、対話音声の認識においては n-best 音声認識候補や word-lattice に対し構文や意味、文脈の制約を用いて再評価する手法が試みられている [6, 8, 3, 7]。いずれも音声認識結果を一意に確定せず、いくつかの候補を言語的な制約を用いて再評価することにより音声認識精度の向上を実現している。

中野らの ISSS は単一の音声認識結果に対する枠組であり音声認識精度の低下に影響を受けやすい。そこで、本研究では n-best 音声認識候補と ISSS を組み合わせることにより対話音声に対する音声理解精度の向上を試みる。提案手法は n-best 音声認識候補を入力とし、それぞれの候補に対する構文解析および理解状態変更規則の適用に関する曖昧性を保持しつつ、逐次的に音声理解を進める。理解状態の曖昧性解消には音声認識、構文解析、理解状態変更の全ての段階における優先度を統合した優先度を用い、beam-search と組み合わせて実用的な処理速度を実現する。

2 n-best 認識候補を用いた逐次発話理解

2.1 逐次発話理解

ISSS について簡単に述べる。ISSS は話しことばを入力とし対話システムなどの理解状態を更新するための音声理解法である。入力文節毎に理解結果を得られるため、音声対話システムに適用すると相槌を含め適切な応答をリアルタイムに出力できる。ISSS は優先度付の構文解析規則および理解状態変更規則を用いる。理解の進行状態は context と呼ばれるフレーム、push-down スタックおよびプライオリティの三つ組みで表現する。フレームとは、例えばユーザ要求や対話進行などを保持する為に用いられる音声対話の内部状態である [2]。スタックは逐次的に入力される単語列の素性構造を保持する。構文解析規則および理解状態変更規則は shift-reduce パージングによりスタックに適用される。shift-reduce もしくは reduce-reduce 操作の曖昧性が発生した場合、context を複製しそれぞれに異なる操作を行う。また、理解状態変更規則により reduce 操作を行う場合、フレーム内容を変更すると同時にスタックを空にする。プライオリティは reduce 操作に対応づけられた個々のスコアを反映し、フレームの優先度を表す。

ISSS はスタックを空にするタイミングが発話の区切りに依存しないため、発話区間をまたがってスタックが保持される場合がある。発話区間をまたがった言語処理とは、このようなスタックに対する構文解析規則および理解状態変更規則の適用である。また、単語列が入力されるたびに最も高い優先度を持つ context を選択することにより、単語列が入力された時点での理解結果を一意に決定できる。

2.2 n-best 認識候補の適用

ISSS は言語解析における曖昧性を context の複製により表現する。本研究では更に優先度付きの n-best 音声認識候補にも対応して context を複製し、音響、言語、発話内／発話間の構文解析および理解状態変更規則に関する優先度を統合して context の優先度を計算する。

2.2.1 アルゴリズム

提案法では、ISSS と同様に音声理解の状態をコンテキスト c で表現する。c は push-down 型スタック st、フレーム f、優先度 $priority$ からなる。すなわち

$$c =< st, f, priority > \quad (1)$$

ユーザ発話から切り出される発話区間に対応する n-best 音声認識結果 NB は優先度 $score$ と認識候補 $wseq$ の組の集合であり、以下の式で表現される。$wseq$ は単語 w の列である。

$$NB = (nb_1, nb_2, nb_3, \ldots, nb_n) \quad (2)$$

1. 時刻 $t = 0$ とし、空のスタック st_0、対話の初期状態 f_0、スコア 0 からなるコンテキスト c_{ini} [式 (1)] を作成して、c_{ini} だけからなる集合 C_0 を作成する。
2. $t = t + 1$ とし、n-best 音声認識結果 NB_t [式 (2)] を受け取る
3. 空集合 C'_{t-1} を作成し、C_{t-1} に含まれる全ての c_i と NB_t に含まれる全ての認識候補 nb_j [式 (3)] の組み合わせについて c_i を複製し c_k として、以下の処理を行う
   (a) スタック st_k に $wseq_j$ を shift する。
   (b) $priority_k = priority_k + \alpha * score_j$ とする。
   (c) C'_{t-1} に c_k を追加する
4. 空集合 C''_{t-1} を作成し、C'_{t-1} に含まれる全ての c_i と Ω に含まれる全ての構文規則 r_j について、以下の処理を行う
   (a) C''_{t-1} に c_i を加える
   (b) st_i に r_j が適用できるならば、c_i を複製し c_k として、以下の処理を行う
      i. r_j により st_k に reduce 操作を行う
      ii. $priority_k = priority_k + \beta * S(r_j)$ とする
      iii. C''_{t-1} に c_k を追加する
5. 空集合 C'''_{t-1} を作成し、C''_{t-1} に含まれる全ての c_i と Ψ に含まれる全ての構文規則 s_j について、以下の処理を行う
   (a) C'''_{t-1} に c_i を加える
   (b) st_i に s_j が適用できるならば、c_i を複製し c_k として、以下の処理を行う
      i. s_j により f_k を書き換える
      ii. st_k を空にする
      iii. $priority_k = priority_k + \gamma * S(s_j)$ とする
      iv. C'''_{t-1} に c_k を追加する
6. $C_t = C'''_{t-1}$ とする
7. C_t の中で、最も $priority$ の高い c を選び、対応する f を時刻 t における音声理解結果とする
8. C_t に含まれる c を $priority$ に関して枝刈りし、c の数を一定数以下に押さえる
9. 手順 2 に戻る

図 1: 逐次発話理解のアルゴリズム

$$nb = < score, wseq > \quad (3)$$

$$wseq = w_1 w_2 w_3 \ldots \quad (4)$$

また、単語列を構文解析する規則の集合を $\Omega = (r_1, r_2, r_3, \ldots)$ で、現在までの理解状態と構文解析された単語列に対してフレーム書換え命令を対応づける理解状態変更規則の集合を $\Psi = (s_1, s_2, s_3, \ldots)$ とする。r、s それぞれには優先度が対応づけられており、$S(r)$、$S(s)$ で参照できるものとする。r、s は shift-reduce parsing における reduce 操作を行う。

以上の要素を用いて、n-best 認識結果を用いた逐次発話理解は図 1 の手順で行われる。

3 評価法

本章では提案法の評価方法について述べる。一般に WER(word error rate) が音声認識の性能尺度として用いられる。しかし音声対話システムなどにおいては、WER の大小よりもむしろ入力発話の持つ意味を正確に理解すること、すなわち理解状態に反映させることが重要である。そこで、本研究では提案法を音声対話システムに適用した際の理解フレーム内容に基づく音声理解率を評価基準とする。

ところで、提案法は複数の発話区間をまたがって言語処理を行うこと、また理解状態更新規則は現在までの理解状態すなわち文脈に依存して入力の解釈を行うことから、音声認識のように一つの発話区間だけを対象として認識や評価を行っても提案法の効果は正確に評価されない。すなわち、処理の対象には一つの発話区間だけではなく、文脈情報を含める必要がある。

そこで本研究では、人間—音声対話システム間の対話コーパスを用いて提案法を評価する。対話コーパスにはシステム発話、ユーザ発話およびシステムが主導権を取った時点の理解フレーム内容からなるものを用い、コーパス中にあらわれるユーザ発話毎に局所的な対話をシミュレートして理解結果を得る。具体的には、当該ユーザ発話以前に最後にシステムが主導権を取った時点の理解フレーム内容を対話の初期状態として、それ以降当該発話までにコーパス上にあらわれるシステムおよびユーザからの発話を全て理解システムに入力した結果として得られる理解フレーム内容を音声理解結果とする (図 2 の上半分参照)。この評価方法は、ユーザはシステムが最後に主導権を取ってから出力したシステム応答にのみ依存して発話を行うとの仮定により成立する[1]。

3.1 データ収集

評価対象として、人間—音声対話システム間の対話コーパスを収集した。データ収集のための音声対話システムは当研究所で開発した音声対話システムツールキット WIT [5] を用いて作成した。タスクは会議室予約であり、一つまたは二つの会議室と開始

[1]本来、ユーザは対話開始時からの全ての発話履歴に依存して発話すると見なすのが妥当である。しかしコーパス中のユーザ発話はコーパスを収集した時のシステム発話にも大きく影響されるため、ユーザ発話のみを対象にした対話全体の再評価は不適切である。このため近似を用いた。

時刻と終了時刻と一つまたは二つの使用日を音声で入力する。対話の進行状態は8個のスロットからなるフレーム表現 [2] で記述される。8個のうち6個はドメイン依存でありユーザ要求に含まれる予約対象を保持する。残りの2個は対話進行の状態を保持するものであり、未確認／確認済フラグおよび対話主導権の所有者を保持する。語彙は171語であり辞書は単語エントリに対して単語カテゴリ素性、意味素性および格素性を与える。23個の構文解析規則と33個の理解状態変更規則を用いた。データ収集は簡易防音を施した実験ブースにて、音声対話システムとの対話経験の無い被験者を対象に行った。各被験者とも収録は18対話行い最初の2対話を除いた16対話をデータとした。システムの発話内容、発話開始、終了時刻、ユーザ発話の開始時刻、終了時刻およびシステムが発話を開始する時点でのフレーム内容、および実際のユーザ発話音声とシステム発話音声がそれぞれ1チャネルづつ記録された。対話収録後全てのユーザ発話は書き起され、目視による発話区間検出誤りの修正を行った。得られた対話データを表1に示す。

収録音声	2ch, 16bit, 16kHz
話者	男性6名、女性10名
総対話数	256対話
総ユーザ発話	3489発話区間

表 1: 収録データ

3.2 評価尺度

理解精度の評価は、システムが主導権を取った時点のフレーム内容と、正解フレームおよび評価実験により得られる理解フレーム候補との差分の比較に基づく。各スロット毎に理解フレーム候補の差分と正解差分と比較し、表2に示す五種類にラベル付けする (図2参照)。

ラベル	名前	説明
D	脱落誤り	正解が変化し、理解候補は無変化
I	挿入誤り	正解が無変化、理解候補は変化
S	置換誤り	正解、理解候補共に変化したが値が異なる
CU	更新正解	正解、理解候補共に変化し値も等しい
CL	非更新正解	正解、理解候補共に無変化

表 2: labelling scheme

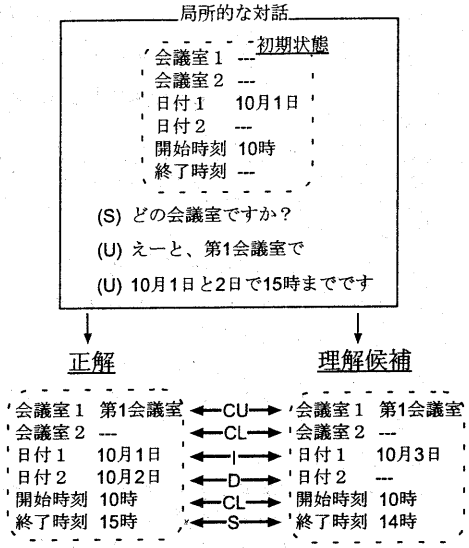

図 2: labelling for evaluation

それぞれのラベルの総数から以下の評価尺度を求める。

スロット正解率 (slot correct rate)

$$\frac{\text{正解と同じ結果のスロット数}}{\text{総スロット数}} = \frac{CU+CL}{\text{総スロット数}} \quad (5)$$

スロット更新精度 (update precision)

$$\frac{\text{候補が正しく変化したスロット数}}{\text{候補が変化したスロット数}} = \frac{CU}{CU+S+I} \quad (6)$$

スロット挿入誤り率 (insertion error rate)

$$\frac{\text{候補が変化したスロット数}}{\text{正解が無変化なスロット数}} = \frac{I}{CL+I} \quad (7)$$

スロット脱落誤り率 (deletion error rate)

$$\frac{\text{候補が無変化なスロット数}}{\text{正解が変化したスロット数}} = \frac{D}{CU+S+D} \quad (8)$$

スロット置換誤り率
(substitution error rate)

$$\frac{\text{候補が誤って変化したスロット数}}{\text{正解が変化したスロット数}} = \frac{S}{CU+S+D} \quad (9)$$

また、正解と理解候補が全てのスロットで一致する発話区間の割合を音声理解率とする。

音声理解率
(speech understanding rate)

$$\frac{\text{全てのスロットが正しい発話区間数}}{\text{全ての発話区間数}} \quad (10)$$

4 実験

4.1 実験システム

3.1 節のコーパスを評価するため、コーパス収集に用いられたシステムの語彙、構文規則等に追加修正を行って新たに評価システムを作成した。修正された語彙は収録データのうち言い淀みなどを除いた全てのデータをカバーし、構文解析規則はデータの97.65%をカバーする。語彙数は 308 で、36 個の構文解析規則と 35 個の理解状態変更規則を用いた。音声認識には連続音声認識コンソーシアム 2000 年度版ソフトウェアである julius3.2[1]、および付属の音響モデルを用いた。言語モデルはコーパスを 8 分割し held-out 法により作成した。tri-gram 言語モデルの平均テストセット・パープレキシティは 3.64 であった。コーパスに含まれる各ユーザ発話に対し上位 10-best 候補を求めた。単語認識精度は 82.23% であり、文正解率は n=1,3,5 に対し 53.28%、63.08%、65.32%であった。また、各候補について音声フレーム毎に認識結果の viterbi path 上における状態と最も高いスコアを与える状態との対数出力尤度比を求め、フレーム長で正規化して発話の音声認識スコアとした。図 1 における α、β および γ は全て 1.0 とした。

4.2 正解データ

本研究では対話システムの理解フレーム内容を評価対象とするため、正解フレームを求めるためには人手による理解フレーム内容のラベリング作業が必要となる。しかし、リソース上の制約やラベラー間の不一致の問題などが想定されるため、書き起こしを 1-Best 認識候補として提案手法による評価を行った結果を正解フレームの近似的とした。なお予備実験として評価データの一部に対し研究者による正解タグ付けを行い近似値を評価した結果を表 3 に示す。スロット更新精度 98.27%、音声理解率 96.91% であり、この近似は十分正確であると考える。

更新精度	正解率	挿入誤	脱落誤	置換誤
98.27%	99.33%	0.33%	1.48%	0.49%

音声理解率
96.91%

表 3: 近似正解値の評価

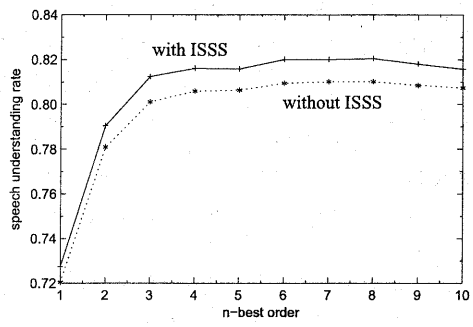

図 3: speech understanding rate

4.3 結果

図 3 に n-best 次数毎の音声理解率を、ISSS を用いる場合と用いない場合にわけて示す[2]。n-best 次数を上げるにつれ音声理解率の向上が認められるがその効果は $n = 3$ 以降で飽和する。また、全ての次数にわたって ISSS が優れた結果を示している。これは評価対象コーパスに複数の短い発話区間にまたがって意味をなす発話が少なからず出現し、ISSS がそれらに有効であったことを示している。なお $n = 3$ における音声理解率は 81.23% であり、$n = 1$ の場合 (72.74%) に比べ絶対値で 8.49% の向上である。

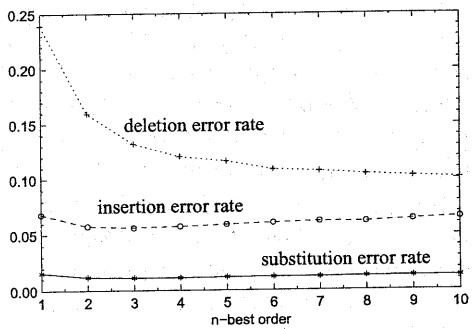

図 4: insertion, deletion and substitution error rate with ISSS

次に、図 4 にスロット毎の挿入、脱落および置換誤り率を示す。n-best 候補の導入により脱落誤りが大幅に減少するが、置換および挿入誤りがわずかに増加する。脱落誤りの減少は上位の認識候補に正解が含まれず下位に含まれている場合には提案方式が効果的であることを示している。挿入誤りは $n = 4$、

[2]ISSS を用いない場合とは、図 1 の手順 2. において n-best 音声認識結果を受け取る際に C_{t-1} に含まれる全ての c のスタックを空にして実現した。

置換誤りは $n=3$ を境に増加するが、これは下位の音声認識候補に構文規則や理解状態変更規則に適合するものが存在した場合、α、β、γおよび優先度の与え方が適切でない場合が存在したことを示している。ただし n-best 次数を高くした場合でも挿入および置換誤り率の上昇はわずかであり、次のグラフからわかるように脱落誤りの減少による効果と相殺される。

なお、図が繁雑になるため ISSS を用いない場合の結果を略したが、全ての次数にわたって ISSS がわずかにまさる結果を示した。

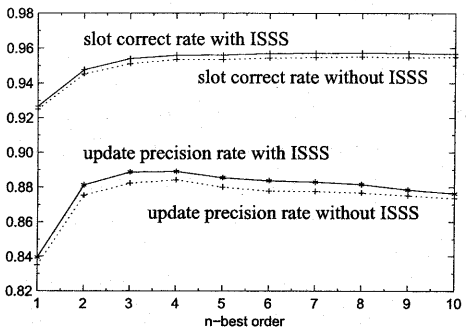

図 5: update precision and slot correct rate

図 5 にスロット毎の更新精度 (precision rate) およびスロット正解率 (correct rate) を示す。スロット更新精度は $n=4$ を境に減少するが、スロット更新精度は脱落誤りを反映しないため $n=4$ 付近からの挿入および置換誤りの増加が直接反映されていると思われる。一方、脱落誤りも反映するスロット正解率は n-best 次数が増加しても減少傾向を示さない。すなわち、$n>4$ の場合にスロット更新精度が低下しても脱落誤りの減少による効果と相殺されることがわかる。また全ての次数にわたって逐次発話理解を用いた場合にわずかながら精度の向上が見られ、逐次発話理解の優位性が確認される。

5 結論

話し言葉に見られるような複数の音声区間にまたがって意味をなす入力発話に対応しながら、より高い音声理解精度を得るため、n-best 音声認識結果を用いた逐次発話理解方式を提案した。提案方式は音声認識における曖昧性を n-best 認識結果の形で保持し、構文解析および理解状態変更規則適用における曖昧性と合わせて音声理解文脈の多重化を行うことにより、発話が入力されるたびに音声理解結果を決定しつつ、発話をまたがる規則適用を可能とする。

また、音声対話における音声理解の精度を評価するため、音声対話をシミュレートして得られる理解結果を用いた評価尺度を導入した。

プロトタイプシステムと人間との対話コーパスを対象に評価実験を行った結果、n-best 認識結果を用いた逐次発話理解方式は n-best 認識結果を用いない場合に比べ絶対値で最大 8.49% の音声理解率向上を示し、逐次理解方式は全般に渡って従来方式をしのぐ結果を示した。これにより、提案方式の有効性が示された。

今後は、規則に与えられるスコアや、認識、構文、理解の各段階のスコアに対する重み係数の最適化を検討する。

参考文献

[1] http://www.lang.astem.or.jp/CSRC/.

[2] D. G. Bobrow, R. M. Kaplan, M. Kay, D. A. Norman, H. Thompson, and T. Winograd. "GUS, A Frame-Driven Dialog System". *Artificial Intellignece*, Vol. 8, pp. 155–173, 2001.

[3] J. Chappelier, M. Rajman, R. Aragues, and A. Rozenknop. "Lattice parsing for speech recognition". In *Proc.* **TALN**, pp. 95–104, 1999.

[4] Mikio Nakano, Noboru Miyazaki, Jun-ichi Hirasawa, Kohji Dohsaka, and Takeshi Kawabata. "Understanding Unsegmented User Utterances in Real-Time Spoken Dialogue Systems". In *Proc.***ACL**, pp. 200–207, 1999.

[5] Mikio Nakano, Noboru Miyazaki, Norihito Yasuda, Akira Sugiyama, Jun-ichi Hirasawa, Kohji Dohsaka, and Kiyoaki Aikawa. "WIT: A Toolkit for Building Robust and Real-Time Spoken Dialogue Systems". In *Proc. first ACL* **SIGDIAL** *workshop*, pp. 150–159, 2000.

[6] Pieraccini R., E. Tzoukermann, Z. Gorelov, J. Gauvain, E. Levin, C. Lee, and J. Wilpon. "A Speech Understanding System Based on Statistical Representation of Semantics". In *Proc.* **ICASSP**, pp. I–193–I–196, 1992.

[7] Carmen Wai., Roberto Pieraccini, and Helen M. Meng. "A dynamic Semantic Model for Re-scoring Recognition Hypotheses ". In *Proc.* **ICASSP**, 2001.

[8] Wayne Ward and Sheryl Young. "Semantic and pragmatically based re-recognition of spontaneous speech". In *Proc.* **EUROSPEECH**, pp. 2143–2147, 1993.