音声言語情報処理 29－19
(1999. 12. 20)

[特別講演] 音声認識における環境適応技術

松本 弘

信州大学 工学部

〒380-8553 長野市若里４丁目１７番１号
Tel: 026-269-5249 ／ e-mail: matsu@sp.shinshu-u.ac.jp

あらまし

実環境に於ける音声認識において耐環境性は、音声認識を広い分野に応用する上で重要な問題である。本稿では、実環境に於ける頑健な音声認識を実現するための付加雑音と乗算性歪みに対する技術を展望し、その課題について述べる。ここでは、音声認識過程における、分析・特徴抽出部、前処理部、認識部の各部に於ける種々の方法について述べる。

キーワード Robust speech recognition, HMM, Adaptation, Convolutional noise, Spectral subtraction

Robust Speech Recognition Techniques in Real Environments

Hiroshi Matsumoto

Faculty of Engineering, Shinshu University
〒380-8553 4-17-1 Wakasato, Nagano-shi
Tel: 026-269-5249 e-mail: matsu@sp.shinshu-u.ac.jp

Abstract

Robust recognition in real environments is crucial to a wide application of speech recognizers. This paper discusses issues concerned with robust speech recognition in real world and reviews the current techniques to cope with additive and convolutional noises. The paper discusses various methods in the HMM framework in three stages of speech recognition process; analysis and feature extraction, and pattern matching stages.

key words

1 はじめに

近年音声認識技術は大きく進歩し、計算機の高性能化に支えられ、商用の高性能な音声認識ソフトウエアも実用化されている。これは、現在の主流である隠れマルコフモデル（HMM）による音響モデルと統計的言語モデルの高精度化並びに探索手法の進歩によるところがおおきい。しかし、統計的手法の欠点として、学習時と認識時の収音環境が異なると認識精度が著しく低下するという問題がある。そのため、実環境における種々の変動要因による認識性能の低下をいかに改善するかが実用上の重要な課題となっている。

実環境に於ける音声認識性能に影響を与える要因は次のように分類される。

- 話者
  - ・ロンバート効果
- 空間伝送系
  - ・環境騒音
  - ・残響、反射音
  - ・話者とマイクロホンの相対位置の変動
- 電気的伝送系
  - ・マイクロホンの特性
  - ・伝送路歪（電話系）
  - ・電気的雑音
  - ・エコー

これらの音声認識への影響は、ロンバート効果を除くと、加法性の雑音、乗算性歪み、非線形歪みに分類される。

これまで耐環境手法については、過去十数年間多くの研究がなされ、大きな進歩も見られるが、まだまだ耐環境性能は不十分である。そこで本稿では、これまでに提案されている雑音と乗算性歪みに対する耐環境手法のうち、主にHMMによる認識で使用される手法について整理し、その課題について述べる。なお、空間伝送系の乗算性歪みへの対策はハンズフリー音声認識における重要な部分であるが、本稿では割愛する。

2 ロバストパラメータ

現在、音声認識に於ける特徴パラメータとして、線形予測分析に基づくLPCメルケプストラム係数とメルフィルターバンク分析に基づくメル周波数ケプストラム係数（MFCC）が良く用いられている[3]。この他、聴覚特性を取り入れたフィルタバンクベースの線形予測分析であるPLP分析は通常のLPC分析に比べ認識性能、耐雑音性に優れていることが知られている[9]。また最近、メル周波数軸上の線形予測分析を時間軸上で処理する方法が考案され、通常のLPC分析に比べ認識性能、耐雑音性が優れていることが示されている[29]。以上の分析法で得られるケプストラム係数は、音声の対数スペクトル包絡を表すパラメータであり、付加雑音や乗算性歪みの影響を受け易いという欠点がある。これに対して、音声と雑音の相関関数の性質の違いを利用して、雑音成分を低減したスペクトル包絡を推定する方法が提案されている。例えば、全極モデルに基づく方法や[]、フィルタ群分析に基づくSBCOR分析法がある[6]。

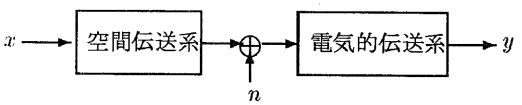

図1 音声認識における環境変動のモデル

以上はスペクトル包絡を表すパラメータであるが、ホルマント周波数またはそれに近いパラメータを抽出することができれば、人間の聴覚と同様耐環境性に優れた認識が可能になるものと期待される。ホルマント周波数を自動的に安定してトッラッキングすることは非常に難しいが、ホルマントに密接に関連したサブバンドのスペクトルピークから、その位置と動きを明示的にパラメータ化し、音声認識に利用した試みがなされている。それによれば、音声バブル雑音下で、MFCCに比べ認識誤りを1／4に低減できることが示されている[1]。従って、ホルマントまたはそれに準じたパラメータを自動的に安定して抽出する手法を確立することは重要である。

3 フロントエンドにおける補償法

3.1 スペクトルサブトラクション（SS）

この方法は雑音の影響を低減する最も単純かつ効果的な方法であり、ＦＦＴに基づくフロントエンドに適している。この方法では、音声と雑音が統計的に独立で、雑音が定常的であると仮定すると、雑音付加音声のパワースペクトル$Y(f)$は音声のパワースペクトル$X(f)$と雑音のパワースペクトル$\widehat{N}(f)$の和となることから、音声パワースペクトル$\widehat{X}(f)$を次式で推定する[14]。

$$\widehat{X}(f) = \max\{Y(f) - \alpha\widehat{N}(f), N_f(f)\} \qquad (1)$$

ここで、αはサブトラクション係数（Overestimation factor）、$N_f(f)$はフロアリングのパワースペクトルである[15]。これより、SS法はWiener Filterの特殊な場合と考えることができる[17]。通常αは1〜2が良い結果を与える。またαを各周波数のSNRに依存した値に設定したSSを非線形サブトラクション（NSS）とよび、通常のSSより優れていることが明らかにされている[16]。フロアリング$N_f(f)$として学習データと同じ雑音スペクトルに設定した場合、結果的に次に述べるマスキング法と同じになる。短時間スペクトル分析では、雑音スペクトルと音声スペクトルは無相関ではない。そのため、低SNRではこれらの相互積の項の影響を無視できなくなるため、サブトラクション後のスペクトルを時間方向に平滑化する方法が提案されている。

3.2 マスキング法

マスキング法は、スペクトルサブトラクションとは逆に、テストスペクトル系列とテンプレートに雑音を付加し、学習時と認識時の不整合を解決する方法である。始めは、BPFスペクトルを特徴量とするDTWに基づく音声認識に適用され、付加する雑音スペクトルは、学習時と認識時の最大の値に設定された[18]。同様の手法がHMMにも適用されている[19]。また、メルフィルタバンクの各チャンネルにマスク信号を付加し、SNR（またはダイナミックレンジ）を目標値に正規化する方法が提案され、雑音レベルだけでなく乗算性歪みの影響を低減できることが示されている[24]。

3.3 ケプストラムバイアスの補償法

3.3.1 ブラインド補償法

CMS（Cepstral Mean Subtraction）[7]は、学習・テスト音声共に、各フレームのケプストラムから長時間平均ケプストラムを差し引くことで、乗算性歪みの差を除去する方法である。これは、乗算性歪みがケプストラム領域で、音声のケプストラムとの和になることを利用したものである。通常、平均化区間は音韻系列の影響を受けないようにある程度の長さが必要であるため、実時間処理に向かないという欠点がある。これに対処するために、状態に依存した平均ベクトルを利用したり[27]、平均の計算にMAP推定を用いる方法が提案されている[28]。

CMSと同様の方法として、ケプストラムの時系列を帯域通過フィルタに通す方法も知られている[12][13]。また、PLP分析の過程で対数スペクトルを帯域通過フィルタに通すLASTA（Relative Spectra）[10]は、CMSと異なり平均化の時定数が160ms程度であるため、音素等に対応するセグメントの変化を強調する効果もある。更に、対数変換を$ln(1+Jx)$に変えたJ-LASTA（Lin-Log RASTA）が提案され、乗算性歪みだけでなく付加雑音の低減にも効果のあることが報告されている[11]。

3.3.2 モデルに基づく補償法

3.3.3 CDCN,VTS,VPS

CDCN（Codeword-Dependent Cepstral Normalization）[20]は、雑音と乗算性歪みを同時に補償しようとする方法である。この方法では、雑音と乗算性歪みを受けた音声のケプストラム$\boldsymbol{y}$を、

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})$$

と音声のケプストラム$\boldsymbol{x}$と補償項$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})$の和で表し、雑音と乗算性歪みのケプストラム係数$\boldsymbol{n}$、$\boldsymbol{h}$をML推定し、事後確率$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{h})$から$\boldsymbol{x}$の最小自乗推定（MMSE）を行う。このとき、音声のグローバルな確率分布として混合ガウス分布を仮定し、発話単位でML推定を行なう。補償項は雑音と加法性歪みに関して非線形な関数であるため、低SNRでの効果に限界がある。その後CDCNは、$\boldsymbol{f}$をTayler級数の0次または1次項で近似したVTS法[21]や、多項式近似を行ったVPS法[22]へと改良され、2チャンネル収音による方法と同程度の効果が得られている。

3.3.4 ケプストラムバイアスのML推定

CDCNと同様にケプストラム領域で、主に乗算性歪みを補償する方法として、ストキャスティックマッチング法[23]とSBR (Signal Bias Removal) [8]がある。これらの方法は、CDCNにおける補償係数（バイアス）自身を未知数として、テスト音声から最尤推定する。SBR法では、音声の確率分布としてCDCNと同様に混合ガウス分布を仮定するのに対し、後者では認識単語系列に対する尤度が最大になるように推定する。ストキャスッティックマッチング法の場合、音声部と無音部とで別のバイアスを推定することにより、CMS法より良い結果が得られている。

4 HMMの環境適応

4.1 HMM合成法

HMM合成法は、ケプストラム係数を特徴パラメータとする無雑音音声のHMMと雑音のHMMから、目的の雑音環境の音声HMMを合成する方法で、PMC (Parallel Model Combination) [35] [36] またはNOVO [38] と呼ばれている。

音声と雑音のパワースペクトルが対数正規分布する場合、これたを重畳して導かれる雑音付加音声の確率分布を、対数正規分布と仮定し、その平均μ_iと共分散σ_{ij}から、次式で対数スペクトル領域の平均μ_i^lと共分散σ_{ij}^lを求め、最後に余弦変換によりケプストラム領域のガウス分布を導く。

$$\mu_i^l = \log\mu_i - \frac{1}{2}\log\{\frac{\sigma_{ii}}{\mu_i^2}+1\} \quad (2)$$

$$\sigma_{ij}^l = \log\{\frac{\sigma_{ii}}{\mu_i\mu_j}+1\} \quad (3)$$

この方法は、HMMを環境に適応させるのに、その環境の雑音モデルを用意すれば良く、学習用の雑音付加音声を必要としない点で優れている。しかし、合成されたHMMは実際の雑音付加音声のHMMよりも認識精度が低い場合がある。この要因としては、(1) 分散が大きいと対数正規分布の近似度が低下する、(2) 短時間スペクトル分析における雑音と音声との相関が考慮されていない、などが考えられ、その改善策も検討されている [25]、[26]。また、乗算性歪みを含めた場合への拡張も行われている [37] [39]

HMM合成法は比較的計算量が多いため、非定常な雑音環境に追随して、迅速にモデルを再合成するのは難しい。そのため高速化法として、DPMC(Data Driven PMC) [40] やFPMC (Fast PMC) [41] が提案されている。またHMM合成時に、雑音モデルの分散を実際の数倍に拡大することで、広範囲のＳＮＲにロバストなモデルを合成する方法も検討されている [29]。この他、HMM合成ではないが、HMMパラメータの雑音成分に関する線形近似により、高速に適応する方法 (Jakobi 法) [42] も低案されている。HMM合成法は非常に有用であり、今後、低SNR における合成精度、非定常雑音に対する頑健性や高速処理に関し更に改良が必要である。

4.2 実データによる適応

4.2.1 MAP 推定法 (直接的適応)

MAP 推定法は、HMMラメータの事前分布を仮定し、学習データが得られたときのパラメータの事後確率 $P(\boldsymbol{\Lambda}|\boldsymbol{O})$ を最大化するように推定する方法で [43]、学習データが増えるに従い最尤推定に漸近する。従って、MAP 推定は、比較的学習データが多いときに有効である。電話音声データベースやクリーン環境のデータベースから自動車内騒音環境のHMMを生成するのに利用され、次のＭＬＬＲと組み合わせることで良い結果が得られている [31]。

4.2.2 写像モデルに基づく環境適応

ある環境のHMMパラメータ Λ_x から別の環境のHMMパラメータΛ_y への変換を写像関数 $F_\nu(\Lambda_y)$ で拘束し、写像パラメータ ν を最尤推定することにより間接的に適応する方法である。平均ベクトルの写像関数としては、話者適応の分野で次の重回帰モデルが良く用いられている。

$$\hat{\boldsymbol{\mu}}_r = \boldsymbol{A}\boldsymbol{\mu}_r + \boldsymbol{b} \qquad (r=1,\cdots,R) \quad (4)$$

重回帰モデルはパラメータが多いので写像に柔軟性があり、付加雑音と乗算性歪みの両方に同時に対応することが可能である。しかし、ロバストな推定を行うためには、学習データ量に応じて推定パラメータ $\{\boldsymbol{A},\boldsymbol{b}\}$ をガウス分布間で共有（結びに）する必要がある。共有クラスタの大きさは、小さいほど音韻等のクラスに依存した適応が可能となるが、重回帰モデルでは、推定パラメータの数が多いため、ある程度以上の大きさがないと正規方程式が不良条件に陥るのでクラスを適切に調整する必要がある。なお、平均ベクトルに加え、共分散行列も変換することで認識精度が若干向上することが確かめられている [48]。更にMLLR の精度を向上させるためには、データ量に応じて共有クラスの大きさと回帰行列の複雑度を適切に制御する必要があると考えられる。これまで不良条件を避ける方法として、回帰行列をブロック対角行列 [48] や帯行列 [44] に制限する方法も試みられているが、特徴ベクトルの離れた成分間の相関が利用できなくなるという欠点がある。

MLLR の適用例として、TI-digit データベースで学習したHMMを、２０人の話者の数字列で走行自動車環境に適応し、整合状態に近い認識性能が得られている [30]。

4.2.3 ストキャスティックマッチング

ストキャスティックマッチング法における写像関数 $G_\eta(\Lambda_x)$ はバイアスモデル $\hat{\boldsymbol{\mu}}_r = \boldsymbol{\mu}_r + \boldsymbol{b}$ で、重回帰行列を単位行列した場合に相当する。しかし特徴空間

の場合と異なり、バイアス b を確率変数（正規分布）と仮定するため、HMMの分散も変換される。[23]。

5 認識時における処理

5.1 MD 法（Missing Data）

この方法は、聴覚による認識を模擬したもので、認識時に特徴パラメータ中の欠損または信頼度の低い成分を除外して尤度計算をする方法 [34] である。欠損成分（Missing Data）の取り扱い方には次の二方法がある。(1) 欠損成分を除いた周辺分布を用いる方法（marginalization）、
(2) 信頼性あるベクトル成分から欠損成分の推定値（MMSE）を求める方法（imputation）。
これらの手法では各特徴ベクトル成分の信頼度の判定が必要であるが、ＳＳ処理結果やSNR の正負による判定が試みられている。対数ＢＰＦスペクトルを特徴ベクトルに用いた実験によれば、ＳＳ処理後に部分的周辺分布を用いた認識を行うことで、ＳＳのみの場合に比べ認識精度が向上することが示されている。この方法では、まだ対数スペクトル成分の相関が考慮されていないという問題点あるが、興味あるアプローチである。

5.1.1 マルチバンドＡＳＲ

この方法は、Flecher の "The Independent - Channel Model"に基づくもので、音声をサブバンドに分割し、各帯域毎にHMMを用意し、それらの尤度を統合することで最終認識結果を得る方法である [32] [33]。尤度を統合するときに、雑音に埋もれた信頼度の低い帯域の寄与を小さく抑えることが可能なため、雑音に強い認識が可能であると言われている。この方法の問題点は、複数帯域の尤度をどの時点で、どのように統合するかにある。これまで、統合時点としては音素毎が、また統合法としては、(a) 線形和、(b) ＭＬＰが検討されている。実際に、雑音環境下の音声認識に適用し、ある程度認識性能の改善が確認されている。この方法では、サブバンドの信頼度の判定と尤度統合への反映の仕方が課題になると思われる。

6 むすび

これまでの耐環境技術のうち、音声認識システムのフロントエンド以降の部分に於ける方法を整理し、それらの課題に等について述べた。本稿では、ロンバート効果への対策法や雑音中の音声検出については触れなかったが、これらは実用上重要な課題である。現在の技術で達成できる実環境下での音声認識精度は、聴覚による認識精度と比べるとまだ大きな開きがあり、今後一層の研究を進める必要がある。

参考文献

[1] K.K.Paliwal, "Spectral subband centroid features for speech recognition," Proc. ICASSP'98, pp.617-620.

[2] B.Strope and A.Alwan, "Robust word recognition using threaded spectral peaks," Proc. ICASSP'98, pp.625-628.

[3] S.Davis and P.Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans., Vol.ASSP-28, No.4, pp.357-366, 1980.

[4] H.Matsumoto et al., "An efficient Mel-LPC analysis method for speech recognition," Proc. of International Conference on Spoken Language Processing, Vol.3, pp.1051-1054.

[5] D.Mansour and B.H.Juang, "The short-time modified coherence representation and noisy speech recognition," IEEE Trans. Vol.ASSP-37, pp.795-804, 1989.

[6] S.Kajita and F.Itakura, "Robust speech feature extraction using SBCOR analysis," Proc ICASSP'95, pp.421-424.

[7] B.Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Amer., Vol.55, pp.1304-1312, 1974.

[8] M.G.Rahim and B.H.Juang, "Signal baias removal by maximum likelihood estimation for robust telehphone speech recognition," IEEE Trans. on Speech and Audio Processing, Vol.4, pp.19-30, 1996.

[9] H.Hermansky, "Perceptual linear predictive (PLP) analysis for speech," J. Acoust. Soc. Amer., Vol.87, pp.1738-1752, 1990.

[10] H.Hermansky et al., "Compensation for the effect of the communication channel in auditoy-like analysis of speech (RASTA-PLP)," Proc. Eurospeech'91, pp.1367-1370.

[11] H.Hermansky and N.Morgan, "RASTA processing of speech," IEEE Trans. on Speech Audio Process, Vol.2, pp578-589, 1994.

[12] H.G.Hirsch et al., "Improved speech recognition using high-pass filtering of subband envelopes," Proc. Eurospeech '91, pp.413-416.

[13] J.Koehler et al., "Integrating RASTA-PLP into speech recognition," Proc. ICASSP'94, pp.421-424.

[14] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans., Vol.ASSP-27, pp.113-120, 1979.

[15] M.Berouti et al., "Enhancement of speech corrupted by additive noise," Proc ICASP'99, pp.913-916.

[16] P.Lockwood and J.Boudy, "Experiments with a nonlinear sepectral subtracter (NSS), hidden Markov models and the projection, for robust speech recognition in cars," Speech Commun, Vol.11, pp.215-228, 1992.

[17] D.V.Compernolle, "DSP Techniques for speech enhancement," Proc. of Workshop on Speech Processing in Adverse Conditions, pp.21-30, 1992.

[18] D.H.Klatt, "A digital filter-bank for spectral matching," Proc. ICASSP'76, pp.573-576.

[19] A.Varga et al., "Noise Compensation algorithm for use with hidden Markov model based speech recognition," Poroc. ICASSP'88, pp.481-484.

[20] A.Acero and R.M.Stern, "Environmental robustness in automatic speech recognition," Proc. ICASSP'90, pp.849-852.

[21] P.J.Moreno et al., "A vector Taylor series approach for environment-independent speech recognition," Proc. ICASSP'96, pp.733-736.

[22] B.Raj et al., "Cepstral compensation by polynominal approximation for environment-independent speech recogintion," Proc. ICSLP'96.

[23] A.Sankar and C.H.Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. on Speech and Audio, Vol.4, pp.190-202, 1996.

[24] T.Claes and D. van Compemolle, "SNR-Normalization for robust speech recognition," Proc. ICASSP'96, pp.331-334.

[25] J.Hung et al., "Improved robustness for speech recognition under noisy conditions using correlated parallel model combination," Proc. ICASSP'98, pp.553-556.

[26] J.Hung et al., "Improved parallel model combination techniques with split Gaussian mixtures for speech recognition under noisy conditions," Proc. ICASSP'99, #2151.

[27] 黒岩眞吾、他、"最尤状態系列を用いた実時間ケプストラム平均値正規化の検討、"電子情報通信学会論文誌、Vol.J82-D-II, No.3, pp.332-339,1999.

[28] T.Kosaka et al., "Instantaneous environment adaptation techniques based on fast PMC and MAP-CMS methods," Proc. ICASSP'98, pp.789-792.

[29] H.Matsumoto et al., "Robust HMM to variation of noisy environments based on variance expansion of noise models," Proc. Eurospeech'99, pp.2387-2390.

[30] Y.Cong and J.J.Godfrey, "Transforming HMMs for speaker-independent hands-free speech recognition in the car," Proc. ICASSP'99, #1721.

[31] A.Fischer and V.Stahl, "Database and online adaptation for improved speech recognition in car environments," Proc. ICASSP '99, #1449.

[32] H.Bourlard et al., "A new ASR approach based on independent processing and recombination of partial frequency bands," Proc. ICSLP'96, pp.426-429.

[33] H.Hermansky et al., "Towards ASR on partially corrupted speech," Proc. ICSLP'96, pp.1578-1582.

[34] L.Josifovski et al., "State based imputation of missing data for robust speech recognition and speech enhancement," Proc. Eurospeech99, pp.2837-2840.

[35] M.J.F.Gales and S.J.Young, "Improved approach to the hidden Markov model decomposition of speech and noise," Proc. ICASSP'92, pp.233-236.

[36] M.J.F.Gales and S.J.Young, "Robust speech recognition using parallel model combination," IEEE trans. on Speech and Audio Processing, Vol.4, pp.352-359, 1996.

[37] M.J.F.Gales and S.J.Young, "Robust speech recognition in additive and convolutional noise using parallel model combination," Computer Speech and Language, Vol.9, pp.289-307, 1995.

[38] F.Martin et al., "Recognition of noisy speech by composition of hidden Markov models," Proc. Eurospeech'93, pp.1031-1034.

[39] Y.Minami, "Univrsal adaptation method based on HMM composition," Proc. ICA'95, pp.105-108.

[40] M.J.Gales et al., "A fast and flexible implementation of parallel model combination," Proc. ICASSP'95, pp.133-136.

[41] Y.Komori et al., "Fast parallel model combination noise adaptation processing," Proc. Eurospeech'97, pp.1527-1530.

[42] S.Sagayama et al., "Jacobian approach to fast acoustic model adaptation," Proc. ICASSP'97, pp.835-838.

[43] C.H.Lee et al., "A Study on Speaker Adaptation of Continuous Density Hidden Markov Models," IEEE Trans., Vol.ASSP-39, No.4, pp.806-814, 1991.

[44] S.J.Cox and J.S.Bridle, "Unsupervised speaker adaptation by probabilistic spectrum fitting," Proc. ICASSP'89, 1, pp.294-297.

[45] P.Kenny et al., "Speaker adaptation in a large-vocabulary Gaussian HMM recognizer," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.12, pp.17-920, 1990.

[46] C.J.Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," Computer Speech and Language 9, pp.171-186, 1995.

[47] V.V.Digalakis et al., "Speaker adaptation using constrained estimation of Gaussian mixtures," IEEE Trans. on Speech and Audio Processing, Vol.3, pp.357-366, 1995.

[48] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," Computer Speech and Language 10, pp.249-264, 1996.