TR-I-0280

文末ピッチパターン情報を用いた文型の認識

Sentence Type Recognition Using Sentence-Final Pitch Contour

水野 理　　ハラルド シンガー　　嵯峨山 茂樹
Osamu MIZUNO　Harald SINGER　Shigeki SAGAYAMA

1992.11

概要

本研究は、韻律情報の音声認識への応用を目指して、疑問文や平叙文などのイントネーションの大きく異なる文に対して、韻律を用いてそれらの自動分類を予備的に検討したものである。一般に平叙文は文末ピッチパターンが下降傾向にあり、疑問文は上昇傾向にあると考えられる。そこで、文末のピッチパターンの傾向をみるために、ピッチパターンに線形近似を施し、その傾きと文型毎の傾向を調べることにより、文の分類を試みた。結果から、このような手法によって文型を判定するのに有効な情報が取り出せることが示された。

ATR Interpreting Telephony Research Labs.
ATR 自動翻訳電話研究所

目 次

1 はじめに

本研究は韻律情報 (特に基本周波数について) の文型 (ここでは、平叙文であるか疑問文であるか、の区別にこの語を用いる) の認識への応用を目指したものである。韻律情報を用いた文型の認識が可能であれば、平叙文であるか疑問文であるかという情報を活用して、認識誤りを軽減できる。また、文型の認識が可能であれば、意味解析の段階で高速化が図れる。一般に平叙文では、文末のピッチが下降傾向にあり、疑問文では上昇傾向にあると考えられる。文末のピッチパターンの傾向を抽出できれば、文型の認識の可能性がある。そのような傾向を見るために文末の基本周波数パターンに対する線形近似を行なった。その傾きをみて文型の認識の可能性を検討し、このような情報が音声認識に有効であることを確認した。

2 ピッチ抽出法

今回実験に用いたピッチ抽出アルゴリズムは、"a super resolution pitch determination algorithm[1]" をもとにしている。これは、相互相関にもとづくピッチ抽出法である。正確で安定したピッチ周波数が得られるため、抽出後の修正の必要はあまりない。このアルゴリズムの使用に当たっては、条件として音声データが十分 low-passed であることが必要である。実験では、音声データにカットオフ周波数が 600Hz のローパスフィルタをかけた後、ピッチ抽出を行なう。フレーム周期は 5ms で抽出を行なう。

3 ピッチパターンの線形近似

最小自乗法による線形近似は次のようにして行なわれる。

ある時刻 j における、ピッチの時系列 θ_j を考える。この時系列の $-N \leq j \leq N$ 区間での傾向をみるために傾き a、切片 b である、直線によって近似を行なう。これは次のような式を考えればよい。

$$E = \sum_{j=-N}^{N} w_j \mid \theta_j - aj - b \mid^2$$

E を最小にするように a、b をもとめる。これには、

$$\frac{\partial E}{\partial a} = 0, \ \frac{\partial E}{\partial b} = 0$$

を解けばよい。

w_j は重みを表し、ピッチパターンの部分的な強調をした線形近似を行なうことができる。

4 線形近似の重み付け

ピッチパターンの重要な部分を強調した線形近似を考える。これは w_j の値のとりかたによって変えることができる。

今回実験に用いた重みは次のようになっている。

1. 三角窓による重み付け

   1 つのピッチパターンにおいて、その両端の値は不安定であることが多い。両端のピッチの影響を小さくするために w_j の値にパターンの中心を 1 として両端に向かって小さくなるような三角窓を用いる。この三角窓による重みを付けた線形近似は、文末の最後に現れたピッチパターンに対して行なわれる。

2. ノコギリ波形による重み付け

   文末のピッチの傾向を見るために、文末を一定区間 (実験では 500ms) 抜きだしそこに現れたピッチパターン全体を線形近似する。一般に文末にいくにつれてピッチは下降する傾向があるので、疑問文に現れるような文末のわずかな上昇を見るために、文末にいくにつれて重みを強くしていく。そのために、文末として抜き出した区間 T(=500ms) のうち、0ms から順に重みを強くして Tで 1 になるような次の関数を考える。

   $$w_j = \frac{1}{T} j \qquad (1 \leq j \leq T)$$

   関数がノコギリ波形の 1 つであることから、ノコギリ波重みとする。

3. 指数関数による重み付け

   ノコギリ波重みよりさらに文末を強調した重みを付ける関数として指数関数重みを用いる。時定数を τ として T(=500ms) で 1 になるようにする。

   $$w_j = e^{(T-t)/\tau} \qquad (1 \leq t \leq T)$$

   この指数関数重み付けには、非常によい性質がある。仮に、文末付近が有声音でなくピッチが抽出できない場合でも、指数関数はどの断片でも相似であることから、ピッチ抽出された区間の最後から指数関数重みで最小二乗近似しても値は変わらない。(いいかえれば、終端に依存しない。) これは、Tが十分大きい場合であるが、終端の変動幅に比べて十分大きければ現実には問題はないであろう。

また、全ての重み付けに対して、ピッチ抽出の際に求められる相関係数値を信頼度として重みに掛け合わせる。

一般に音声合成では文末の 2〜3 モーラに着目し、疑問文などの抑揚を考えるといわれる。文末として 500ms 抜き出したのは、この 2〜3 モーラを十分含むように設定されたためである。

5 表層 IFT に基づく文型分類

文型の認識の評価のためには、使用する音声資料に対して文型ラベルを与えることが必要である。今回の研究で用いている表層 IFT(surface Illocutionary Force Type)[2] は、人間の対話モデルの発話行為のレベルでの記述を行なったもので、音声データに対応するテキストデータにこのようなラベルが付されているので、これを利用した。

表層 IFT は次のタイプに分かれる。

Type	タイプ	例文
questionif	真偽疑問	割り引きはないのですか
questionref	疑問語疑問	参加料はいくらですか
questionconf	確認	振り込まれておられますね
phatic	挨拶	もしもし
request	依頼	手続きをしてください
inform	情報伝達	こちらは会議事務局です
response	応答	はい
expressive	感情表現	ありがとうございました
promise	約束	登録用紙を送らせていただきます

本報告では、疑問文とは、questionif、questionref、questionconf の 3 つを指し、平叙文とはそれ以外をいうことにした。

なお、各 type に対する日本語訳 (タイプ) は文献 [2] には載っていない。日本語に訳した場合は、若干ニュアンスが異なるため、ここで載せた日本語訳は各 type の意味を全て含むものではない。

6 specview に基づくピッチパターンの表示

ピッチパターンやピッチ抽出誤りの確認を行なうために、ピッチパターンの表示を行なうプログラムを開発した。これにより、ピッチパターンのログ (log) スケールでの表示と各ピッチパターンに対する線形近似の結果を表示した。

図 1 と図 2 に文型認識誤りの無かった文章に対する specview での表示結果の例を示す。図 1 は平叙文 (inform)、図 2 は疑問文 (questionref) である。図 3 と図 4 に文型認識誤りのあった文章に対する specview での表示結果の例を示す。図 3 は平叙文 (inform) において、文末が無声化したために正確なピッチパターンが得られなかった例である。図 4 は疑問文 (questionif) において、文末のピッチパターンが下降している例である。

7 文末ピッチパターンの線形近似実験

7.1 音声資料

今回使用した音声資料は、ATR モデル会話 224 文章で、男性話者 MAU による連続発声したものを使用した。各データには表層 IFT に基ずくラベルが付けてある。表層 IFT によるラベルの付けてあるもの 225 文章である。ただし、話者 MAU においては表層 IFT によるラベルの付けてあるモデル会話は MAU_M51_15 に相当するデータが無かったため、224 文章とした。

224 文章のうち、平叙文は 188 文章、疑問文は 36 文章であった。量子化ビット数 16、サンプリング周波数 12kHz のものを使用した。

7.2 線形近似

重み付けには、先に述べたように

1. 三角窓による重み付け
2. ノコギリ波形重み付け
3. 指数関数重み付け

の3種類を用いた。文末へいくにしたがってピッチパターンが下降していく傾向は、疑問文、平叙文を問わずある。文末全体の線形近似を行なった場合、どの傾きも負になる可能性がある。そこで、指数関数の重み付けによる線形近似の傾きと、ノコギリ波形の重み付けによる線形近似の傾きの差をとる。この結果が正であれば、文末に向かって上昇する傾向があるといえる。また、負であれば、文末に向かって下降する傾向があるといえる。この傾きの正負の値によって疑問文、平叙文の判定を行なってみた。

7.3 実験結果

方式1として三角窓による重み付けの線形近似結果を表1に示す。方式2として指数関数の重み付けによる線形近似の傾きと、ノコギリ波形の重み付けによる線形近似の傾きの差をとった結果を表2に示す。方式3として、2つの結果を合わせた結果を表3に示す。方式1では、疑問文中22文章(61%)について傾きが正となった。平叙文については162文章(86%)について傾きが負となった。傾きが正となった文章の中で疑問文であったものは、22/48 (45%)であった。付録A.1に疑問文において傾きが負になったもの（１６文章）のリストを示す。付録A.2に平叙文において傾きが正になったもの（特に傾斜の強い１０文章）のリストを示す。

表1: 三角窓による重み付けの線形近似結果

	傾き	
	傾き正の値	傾き負の値
疑問文	22	14
平叙文	26	162
total	48	176

方式2では、疑問文中28文章(78%)について傾きが正となった。平叙文については120文章(64%)について傾きが負となった。傾きが正となった文章の中で疑問文であったものは、28/96 (29%)であった。図５、６にリストA.1中の音声ファイルのピッチパターンとノコギリ波形重みによる線形近似の結果（図の破線部分）と指数関数重みによる線形近似の結果（図の実線部分は$\tau = 100$の時の結果、３点鎖線は$\tau = 10$の時の結果）を示す。図７にリストA.2中の音声ファイルのピッチパターンと線形近似の結果を図５、６と同様に示す。

表 2: 指数関数の重み付けによる線形近似の傾きとノコギリ波形の重み付けによる線形近似の傾きの差

	傾きの差	
	傾きの差が正の値	傾きの差が負の値
疑問文	28	8
平叙文	68	120
total	96	128

表3では、2つの結果が共に正であったものは36文章で、そのうち疑問文は20文章(56%)であった。2つの結果が共に負であったものは116文章で、そのうち平叙文は110文章(95%)であった。

表 3: 方式1と方式2を合わせた結果

	傾き			
	共に傾き 正の値	方式1のみが傾き 正の値	方式1のみが傾き 負の値	共に傾き 負の値
疑問文	20	2	8	6
平叙文	16	10	52	110
total	36	12	60	116

8 実験結果への考察

傾きの正負に着目して、文型の分類を行なってみた。以上の結果から次のようなことが分かった。

三角窓による重み付けの線形近似結果

- 疑問語疑問 (questionref) では、全て文末のピッチパターンが上がっておらず、線形近似での傾きが負になった。これは、疑問語にイントネーションが移行しているために文末でのピッチが上がらないのだと考えられる。
- 平叙文中で傾きが正になった文章のうちで、特にその傾斜の強い10文章について調べたところ、8文章については、ピッチ抽出のエラーが生じていた。残りの2文章は“いいえ”(response) という文章であった。response に含まれる文章は、一般的な平叙文とは異なりアクセント型で文全体のピッチパターンが構成されている。“いいえ”のように文の後ろにアクセントか来る場合などは、傾きが正になる可能性がある。

指数関数の重み付けによる線形近似の傾きと、ノコギリ波形の重み付けによる線形近似の傾きの差をとった結果

- 疑問文に対しては結果が向上した。最後のピッチパターンの傾きが負になった文章のうち questionref では 6 文章について改善されている。
- 平叙文については、逆効果となったものが多くなっている。最後のピッチパターンの傾きが正になったもののうち、傾きの差が負になったものが 1 つあるが、ピッチ抽出に誤りがあるものなので、改善されたとはいえない。

方式 1 と方式 2 を合わせた結果

- 今までの 2 つの結果を合わせて、共に正であれはもっとも疑問文らしいと考えられ、共に負であればもっとも平叙文らしいと考えられる。もっとも疑問文らしいとされた文章のうち、56%が疑問文であった。これは、いままでの結果のうちもっとも良いものである。
- 逆にもっとも平叙文らしいとされた文章のうち 95%が平叙文であった。
- 問題点は傾きの正負の混じった 72 文章に対する判定ができないていことである。

9 まとめ

文型の認識のために文末ピッチパターンの線形近似を行なった。この手法は単純ではあるが、良好な認識結果を得られる可能性がある。つまり、文末のピッチパターンの傾向をみることで平叙文と疑問文の認識がある程度可能であるということである。特に、平叙文ではその効果は高く、認識の誤ったものの多くがピッチ抽出のエラーによるものであることから十分改善の余地があるといえる。ピッチ抽出エラーがあるデータを除外すれば、文型認識率はもっと高くなる。

疑問文に対しては疑問語疑問に対する対策を考える必要がある。疑問語にイントネーションが移行しやすいために、文末が上昇傾向にならないためである。これは、音韻認識に依存する必要性があると考えられる。

また、ピッチ抽出法の改善も重要な課題の 1 つである。今回、ピッチ抽出の結果には一切修正を施さなかった。全体の結果をみると、今回採用したピッチ抽出アルゴリズムは倍ピッチ、半ピッチが比較的少ないが、無声化した音声には弱いことである。文末のピッチは無声化し易いため、さらに改善されることが望まれる。

10 今後の課題

今後の課題はつぎのようなものが挙げられる。

1. ピッチ抽出エラー対策
2. 抽出区間の再検討
3. 線形近似の傾き以外の分類

10.1 ピッチ抽出のエラー対策

1. 最小自乗誤差の評価

   ピッチ抽出の結果に倍ピッチや半ピッチがあると最小自乗誤差が大きくなると考えられる。誤差値を評価することで、ピッチ抽出のエラーを検出することができる。

2. 音声区間とピッチパターンの長さ

   音声区間に対してピッチパターンの長さが極めて短ければ、ピッチ抽出のエラーがあると考えられる。音声区間に対するピッチパターンの割合いを評価することで、ピッチ抽出のエラーを検出することができる。

3. 相関係数の総和とピッチパターンの長さ

   ピッチ抽出の際、相関係数がもとめられる。これに閾値を設けピッチとして採用するかを決める。音声区間全体の相関係数の総和に対してピッチパターンとして採用されている相関係数の総和が極めて小さければ、ピッチ抽出のエラーがあると考えられる。2 つの比を評価することで、ピッチ抽出のエラーを検出することができる。

10.2 抽出区間の再検討

1. 500ms の時間幅をさらに小さくする。

   500ms では、その区間に含まれるモーラ数が多くかえって悪い結果を導く可能性がある。さらに縮めた区間での実験が望ましい。

2. ピッチの終端から一定時間音声を抜き出す。

   最後に現れたピッチパターンの終端から一定時間抜き出すことを考える。これによって音素ラベルを用いる必要性が無くなる。

10.3 線形近似の傾き以外の文型分類法

線形近似のみでは、ピッチ抽出の精度や音韻環境に左右されやすい。そこで、ピッチパターンをそのまま利用することを考える。次のような方法が考えられる。

1. ピッチパターンから数ポイント抜き出して比較、分類を行なう。

2. 不連続に現れるピッチパターンに線形補間を施しそこから、数ポイント抜き出して比較、分類を行なう。

3. ピッチの時間変化量 (デルタピッチ) を用いる。

これらに対して十分なデータが得られれば、クラスタリングを行ない、より高度な分類が可能になると考えられる。この問題は、基本的に文型のパターン認識であり、LVQ やニューラルネットなどの識別法が利用できよう。

11 謝辞

研究の機会を与えて頂いたATR自動翻訳電話研究所 および 樽松 明 社長に感謝致します。specviewの改造に関して藤原紳吾さん、ピッチ抽出に関してPaul Bagshawさんから助言を頂いたことに感謝いたします。また、早稲田大学白井克彦教授、小林哲則助教授に感謝致します。

参考文献

[1] Y. Medan, E. Yair, and D. Chazan: "Super resolution pitch determination of speech signals," *Transactions on Acoustics, Speech, and Signal Processing*, ASSP-39(1):40-48, 1991.

[2] M. Nagata: 「統計的な対話モデルの試みとその音声認識への利用」情報処理学会研究グループ資料 92'(P91).

A 付録

A.1 疑問文において最後のピッチパターンに対する線形近似の傾きが負になった文章 (16 文章)

ファイル名	表層 IFT	例文
MAU_M11_05	questionref	“どのようなご用件でしょうか”
MAU_M11_07	questionref	“どのような手続きをすればよいのでしょうか”
MAU_M11_09	questionif	“登録用紙はすでにお持ちでしょうか”
MAU_M21_04	questionref	“いま会議に申し込めば参加料はいくらですか”
MAU_M21_10	questionif	“参加料の割り引きはないのですか”
MAU_M21_13	questionref	“参加料はどのようにお支払いしたらよいのですか”
MAU_M31_08	questionref	“ところで会議での公式言語はなんですか”
MAU_M41_03	questionif	“会議の案内書はお持ちですか”
MAU_M51_06	questionif	“お名前をお伺いできますでしょうか”
MAU_M51_12	questionif	“登録料を払い戻して頂けますか”
MAU_M51_17	questionif	“では誰かが私の代わりに参加することはできますか”
MAU_M71_05	questionref	“何でしょうか”
MAU_M81_05	questionref	“どのような手続きをすればよいのでしょうか”
MAU_M81_11	questionref	“要約はどのような書式で書けばよいのですか”
MAU_M91_04	questionref	“何のご用件でしょうか”
MAU_M91_12	questionref	“では北大路駅からですといくらぐらいかかりますか”

A.2 平叙文において最後のピッチパターンに対する線形近似の傾きが正になった文章 (特に傾斜の強い 10 文章)

ファイル名	表層 IFT	例文
MAU_M11_10	response	“いいえ”
MAU_M21_08	inform	“参加料には予稿集代と歓迎会費が含まれています”
MAU_M21_14	inform	“参加料は銀行振り込みです”
MAU_M31_05	inform	“今回の会議は通訳電話に関連する広範な研究分野を含んでいます”
MAU_M31_13	inform	“英語への同時通訳を用意しております”
MAU_M41_07	inform	“会議は八月二十二日から二十五日まで京都国際会議場で開催され
MAU_M61_05	response	“はい”
MAU_M71_18	response	“いいえ”
MAU_M81_16	inform	“住所は、東京都豊島区東池袋三丁目二番五号です”
MAU_M01_34	inform	“電話番号は三三一の二五二一です”

A.3 specview に組み込んだピッチ周波数パターンの実例

specview にピッチ周波数の表示機能を組み込んで、実際の音声データから得られたピッチ周波数パターンの実例を次ページ以降に示す。直線は、各有声区間での最小二乗直線近似を表す。図1と図2は文型認識誤りの無かった文章に対する specview での表示結果の例である。図1は平叙文 (inform)、図2は疑問文 (questionref) である。図3と図4は文型認識誤りのあった文章に対する specview での表示結果の例である。図3は平叙文 (inform) において、文末が無声化したために正確なピッチパターンが得られなかった例である。図4は疑問文 (questionif) において、文末のピッチパターンが下降している例である。

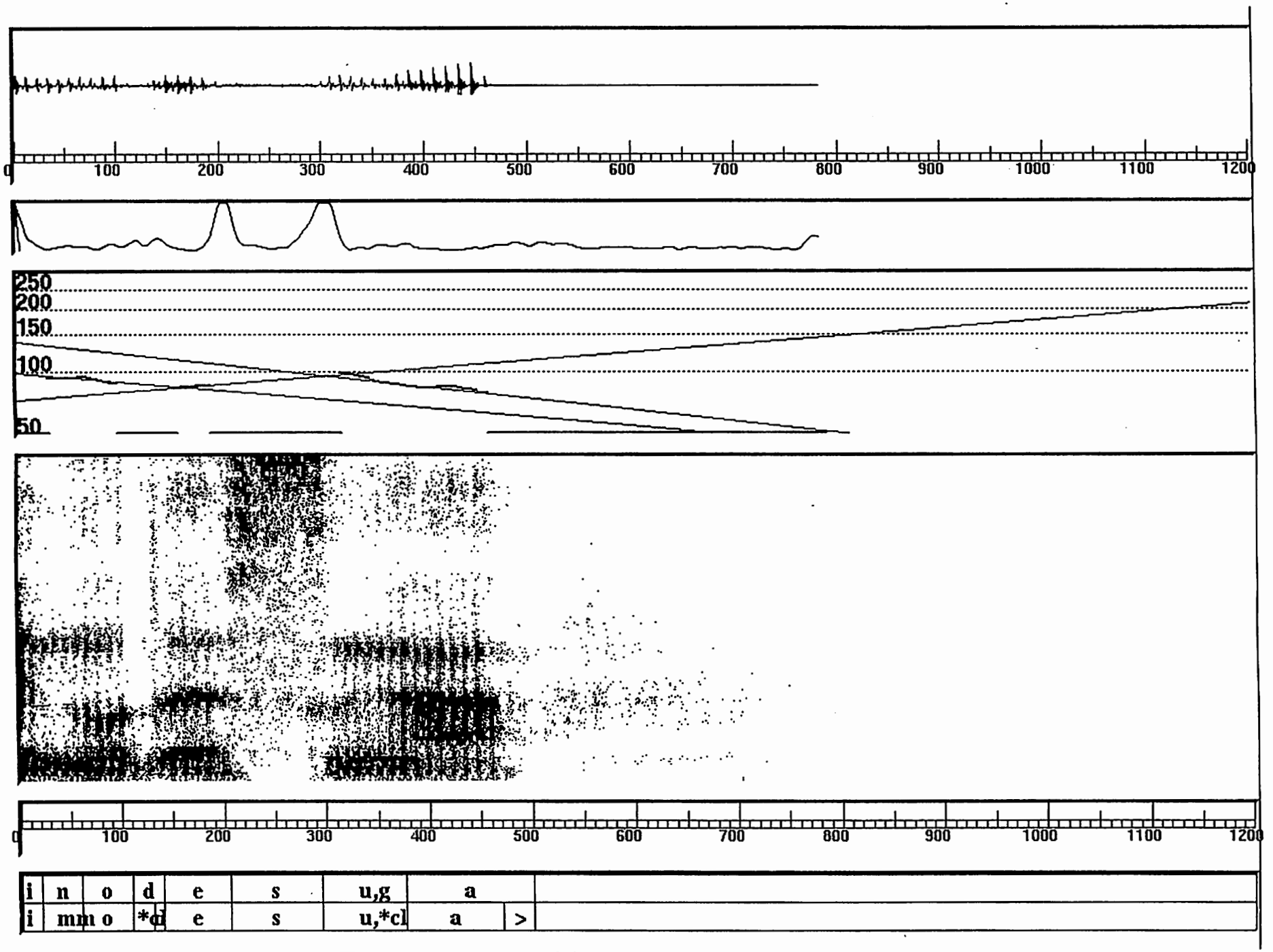

図 1 MAU_M01_04 inform "会議の宿泊施設についてお尋ねしたいのですが"

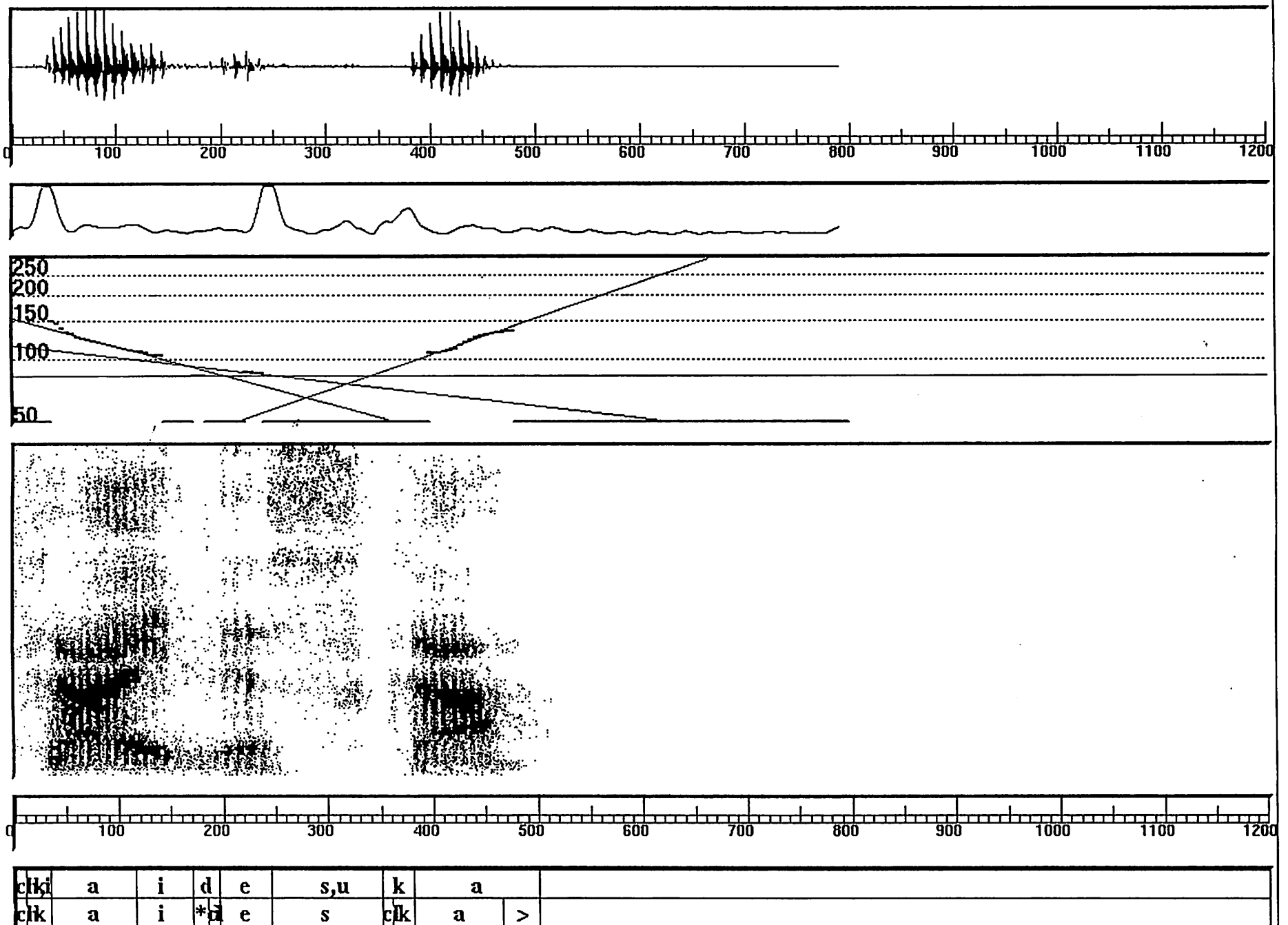

図2 MAU_M01_11 questionref “どちらのホテルが近いのですか”

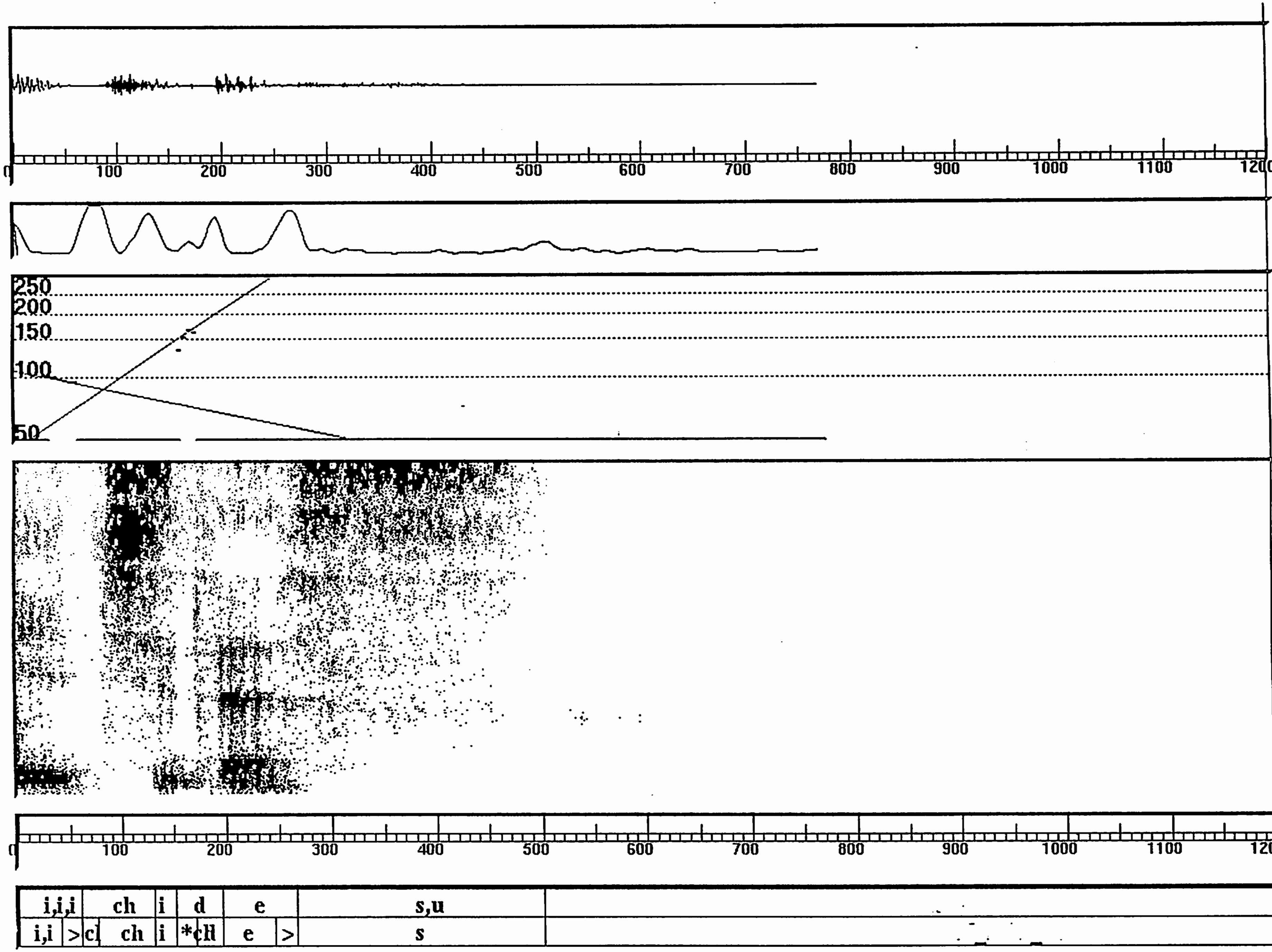

図3 MAU_M01_34 inform "電話番号は三三一の二五二一です"

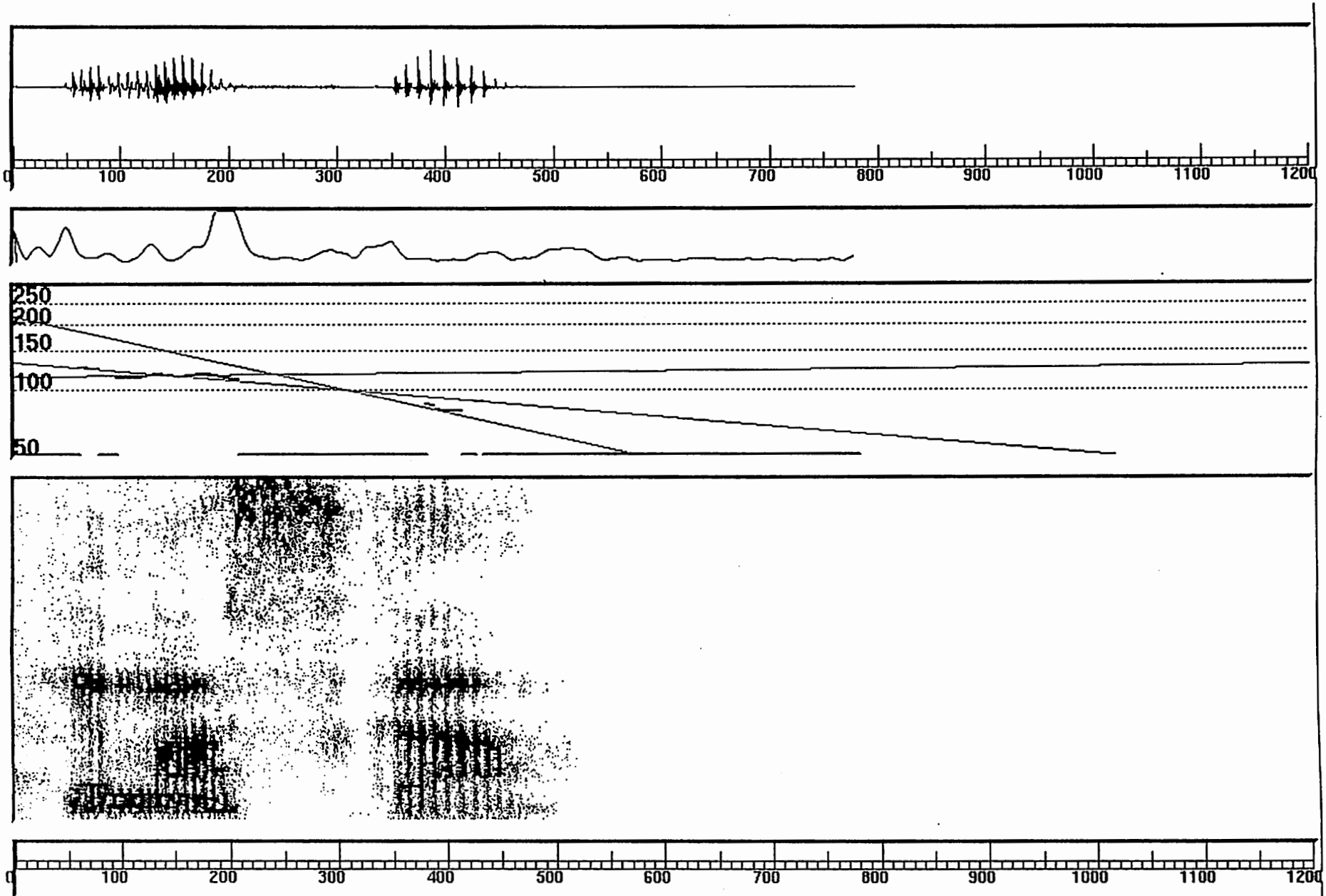

図4 MAU_M51_12 questionif "登録料を払い戻していただけますか"

A.4 方式1において誤りのあった音声のピッチパターンとそれに対するノコギリ波形重み、指数関数重みによるそれぞれの線形近似の結果

方式1において誤りのあった音声のピッチパターンとそれに対するノコギリ波形重み、指数関数重みによるそれぞれの線形近似の結果を次ページ以降に示す。図の太線部分はピッチパターン、破線部分はノコギリ波形重みによる線形近似の結果、実線部分と3点鎖線部分は指数関数重みによる線形近似の結果（それぞれ時定数$\tau = 100$、$\tau = 10$）を示す。図5、6は付録A.1中の16文章について、図7はA.2中の10文章についてそれぞれ結果を示す。

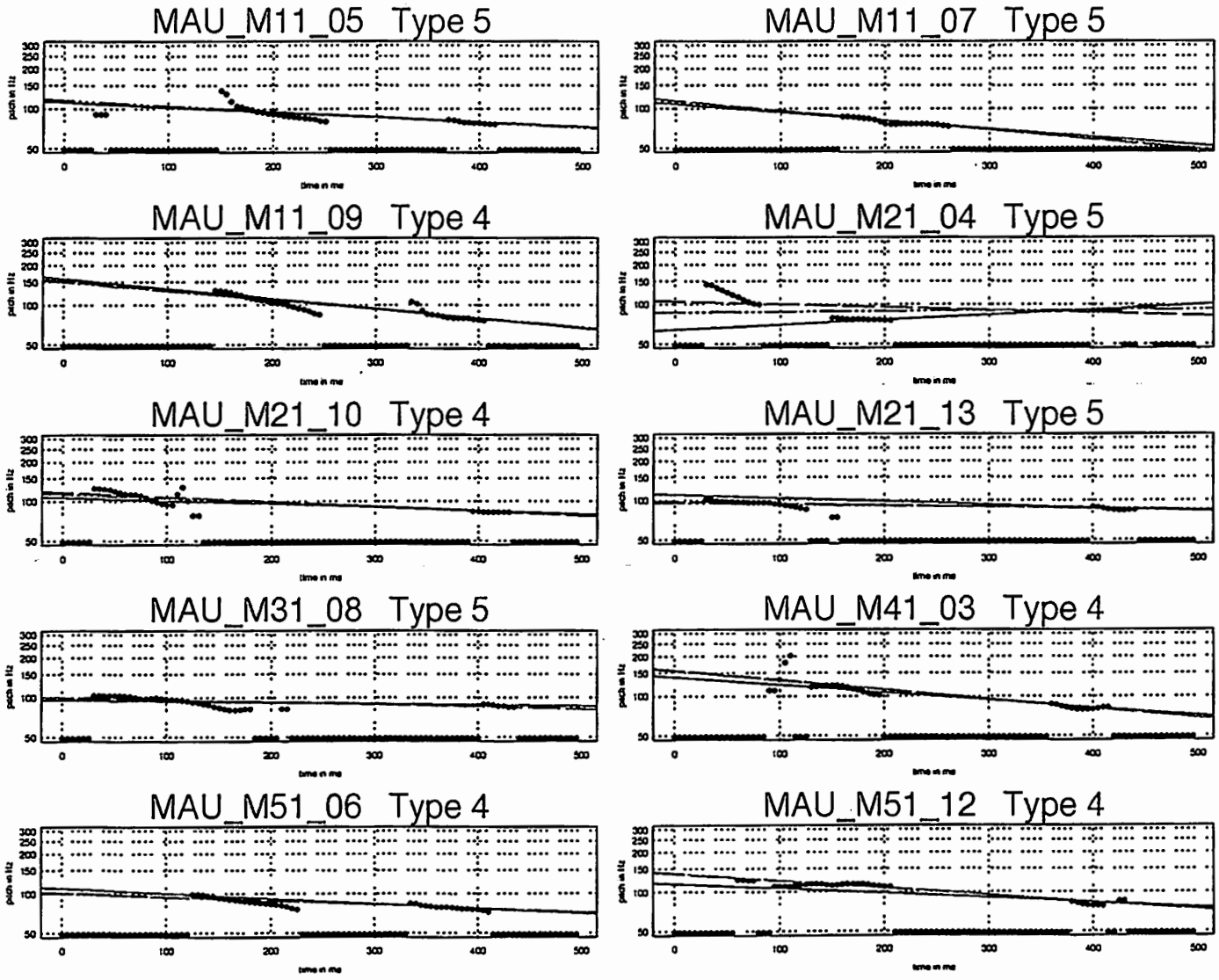

図5 リスト A.1 中の音声ファイルのピッチパターンとノコギリ波形重みによる線形近似の結果（図の破線部分）と指数関数重みによる線形近似の結果（図の実線部分は $\tau = 100$ の時の結果、3 点鎖線は $\tau = 10$ の時の結果）

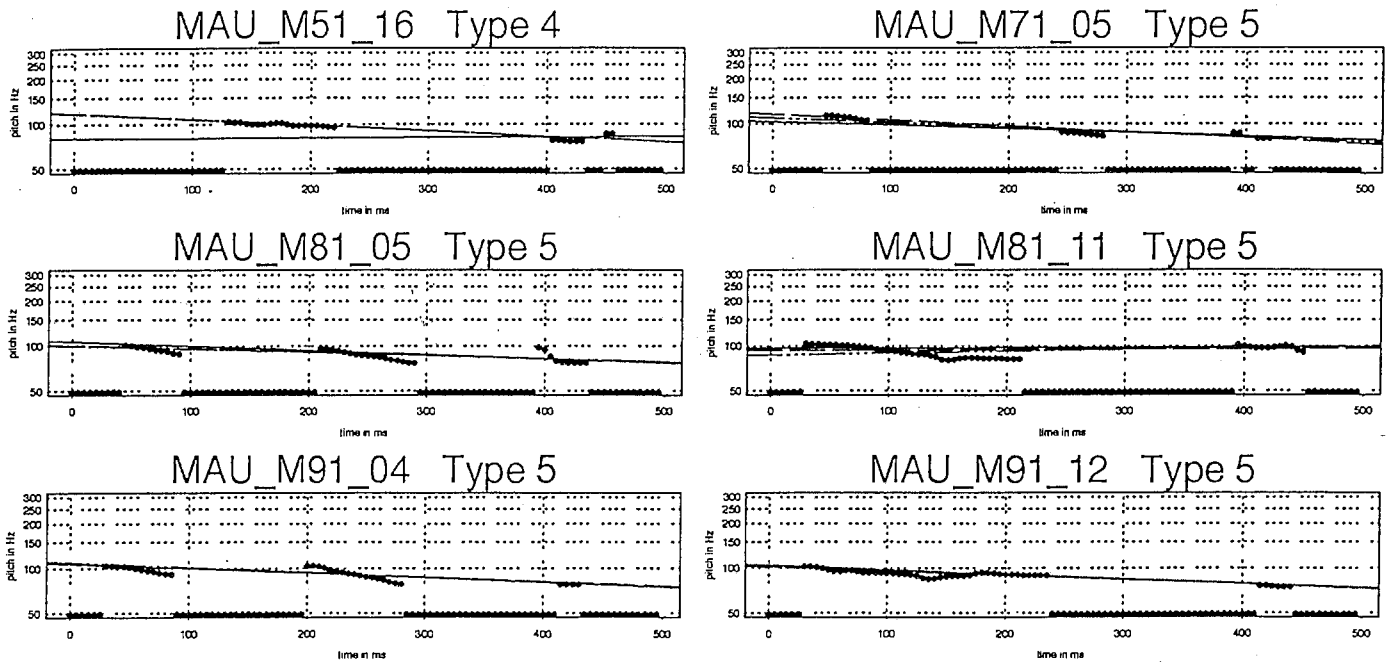

図6 リスト A.1 中の音声ファイルのピッチパターンとノコギリ波形重みによる線形近似の結果（図の破線部分）と指数関数重みによる線形近似の結果（図の実線部分は $\tau = 100$ の時の結果、3 点鎖線は $\tau = 10$ の時の結果）

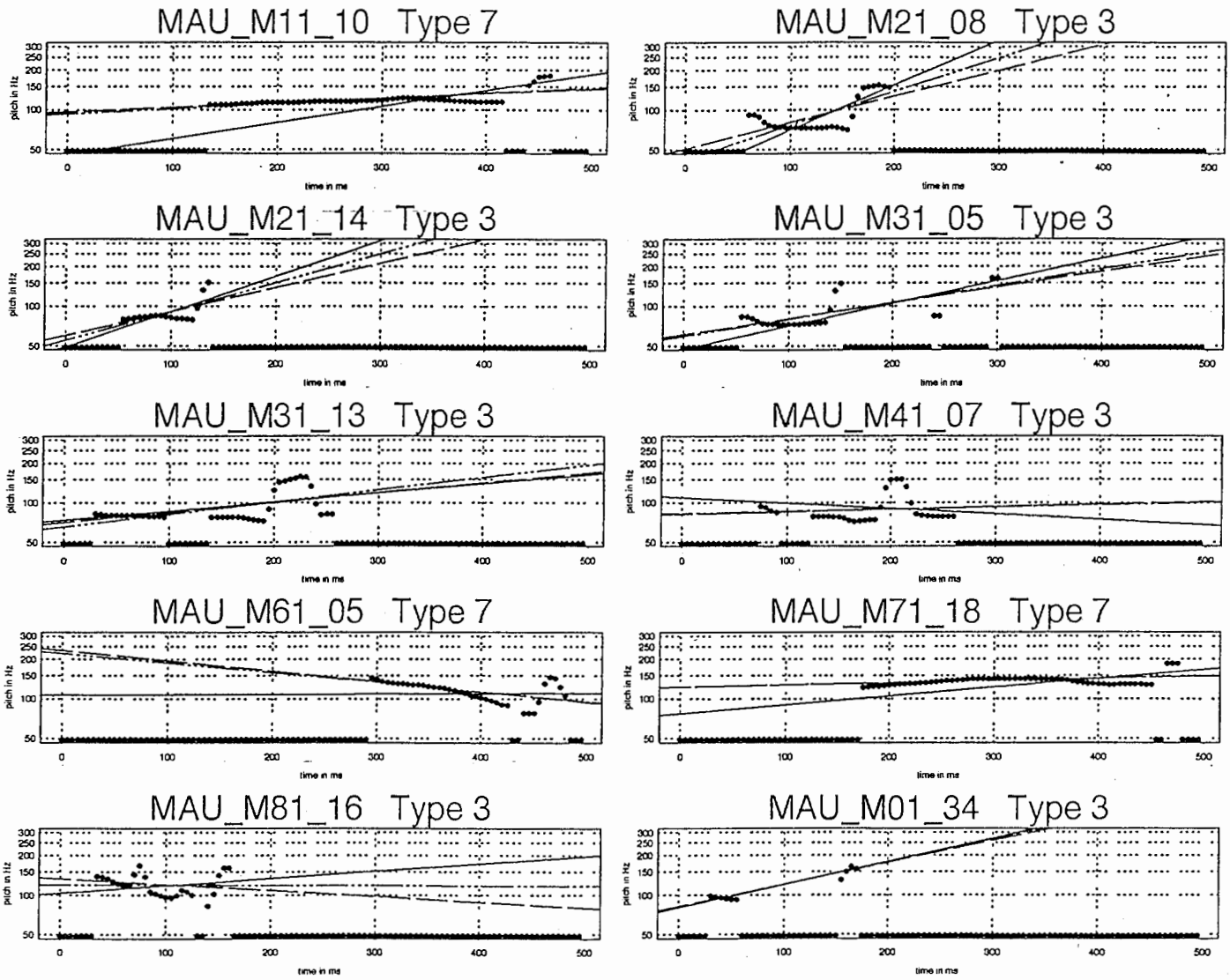

図7 リスト A.2 中の音声ファイルのピッチパターンとノコギリ波形重みによる線形近似の結果（図の破線部分）と指数関数重みによる線形近似の結果（図の実線部分は $\tau = 100$ の時の結果、3 点鎖線は $\tau = 10$ の時の結果）