音声コミュニケーションによる
気持ちのやり取り

ニック　キャンベル

February 14, 2006

1　はじめに

平成 12 年から平成 18 年 3 月までの 5 年間の間、人の話し方コーパスの作成を行った。いく人かのボランティア話者が日常生活をしながら高品質マイクを通して、自分の日常対話の録音を行った [1]。その結果、国際的にも価値ある大規模自然発話音声コーパスができ、すべての発話の書き起こしにより、実際の話し言葉の特徴を分析する際の素材を提供することとなった。コーパスの発話内容は様々であり、丁寧な話の場面ももちろん多いのだが、くだけた話し方のほうが日常対話においては当然ではあるがより普通であることが分かった。本稿ではこの「普通」の話しことばの音声文法構造について述べる。

つまり、よい意味よりもむしろ ”broken” というよくないニュアンスが強い。丁寧な話し言葉の組み立ては、よく考えた上で発せられるものであるため、書き言葉に近い整った構造をとる。一方，日常の多くの場面で見られるくだけた表現はリアルタイムで産出されるもので、話し言葉としてより自然であると考えられるべきである。自然であることは「くだけた」とは必ずしも一致しない。この話し言葉の特別な構造についてどう説明すべきであろうか.

この問題を説明するためには、音声コミュニケーションの役割について考える必要がある。つまり、我々は日常対話において、声のトーンや声質の設定により、心的コミュニケーションを実行している [2]。

2　ノンバーバル音声言語コミュニケーション

『文法と音声 IV』[3]　で既に詳しく述べたように、相手によって声の平均設定が決まる。図 1・図 2 は 1 人話者の 3 年間のデータであり、それぞれの相手による声質とイントネーションのパラメータの変化を示す。上部の NAQ は地声、裏声の次元で変化する声の硬さをはかるメジャーである [4, 5]。NAQ の高い場合、声は柔らかく、ソフトな話し方となる。逆に NAQ の低い場合、声は気息性がなく強い話し方となる。下部は F0（発話音声の基本周波数）で声の平均高さを示す。図 1 は 5 種類の相手、つまり幼児、家族、友達、他人、自分（独り言）を示す。幼児と他人の場合、気遣い度が高く、F0 も NAQ も両方共に高い数値を示す。しかし、独り言の場合、自分への気遣いの必要はなく、F0 が高くても NAQ は低くなる。両図においてすべての差は有意である。

図 1 の家族をより具体的に比較する。図 2 の示した声の特徴で気遣い度という配慮の度合いは話者の母（m 1）・姉（m 5）とも NAQ・F0 とも同一数値を示し、話者との関係が同一であることを示唆する。これに比べて、父（m 2）の F0 も NAQ も高く、さらに幼児（m 3）は F0 も NAQ 顕著に高くなっている。m 4、m 8 の示す人は NAQ が低く、親しさにより気遣いのあまり必要としない関係であることが分かる。m 4 は夫、m 8 は叔母、m 6 は m 5 の 10 歳の息子をそれぞれ示している。韻律情報は主に文構造を示す役割があるが、同時にまた話者の心的情報も伝える役割を持つ。このように声のトーンや声質はパラ言語情報を示す。

＜＜図 1 図 2 ここ＞＞

3　パラ言語情報

声が持つ情報には、言語情報（文内容）はもちろん、非言語情報（話者自身の性別・年齢など）、そしてパラ言語情報（話者意図・態度・感情・発話行為・相手との人間関係等の情報）が含まれている。ある内容を相手に伝える際、言い方によって、伝えるニュアンスは大きく変わる。

講義やニュースを書き起こすと、その文字列からほとんどの意味が伝えられるが、日常の会話音声の書き起こしデータからは、「言い方」の表す意味の違いが含まれていないため、話者意図など文字列だけで表現されない情報が欠落する。例えば、「えー」、「あー」、「あ」、「あっ」、「へー」などの単音であっても話者状況・意図・意見などは示される。語彙的情報そのものよりもこのような単音がもつ情報は韻律や声質の変化が多く、相互的な発話行為の中で、声質のコントロールなどによって微妙なニュアンスの違いを表現できる [6]。

前述の大規模対話コーパスより、1 人話者分の自然対話発話の約６００時間の発話音声を取り出したものは、約１５万発話単位の書き起こしデータとなる。書き言葉の場合は文に簡単に分けることができるが、話し言葉は文の構造をもっていないことが多く、書き起こしデータを単純に文に分けることは困難である。そのため、1 発話は 1 意味単位とのルールで書き起こしを行った。つまり分かり易く言えば、ラベラーに「一行一円」の基準を提案し出来るだけ細かくその音声を

Table 1: Counts of non-verbal utterances in the transcriptions for one speaker in the ESP corpus. Utterances labelled ‘non-lexical’ consist mainly of sound sequences and combinations not found in the dictionary, but may also include common words such as “yeah” “oh”, “uhuh”, etc.

total number of utterances transcribed	148772
number of unique ‘lexical’ utterances	75242
number of ‘non-lexical’ utterances	73480
number of ‘non-lexical’ utterance types	4492
proportion of ‘non-lexical’ utterances	49.4%

Table 2: The most common complete utterances in the corpus, (data from one speaker, numbers show occurrence frequency)。 Note the highly repetitive nature of these common expressions

48038	うん	1733	で	829	ま
15555	あ	1675	ほんで	800	んんん
10961	ふん	1550	うんうん	787	まあ
8408	うーん	1535	もう	751	わかった
7769	え	1428	でも	737	や
5796	ああ	1422	ふんで	730	ありがとう
4891	ほんま	1412	はあ	713	あれ
4610	あー	1370	ええ	703	そうそうそう
3704	んん	1329	そう	692	は
3608	はい	1299	ふんん	692	そうなんや
3374	なんか	1291	ほんまあ	687	あたし
3164	ん	1246	うんうんうん	679	んんーん
3010	いや	1227	あのう	674	はいはい
2942	ふーん	1206	ううん	673	そうそうそうそう
2860	あの	1118	これ	658	フフ
2246	ふうん	1108	そうそう	645	せやな
2238	なあ	1085	おん	623	ほんなら
1871	そうなん	1079	まあな	599	うんうんうんうん
1761	な	903	あああ	588	ほん
1736	うんん	871	だから	583	よいしょ

発話単位に分けるよう指示した。1 イントネーションユニットは 1 単語でも３０単語であってもひとまとまりと考える。

これらの書き起こしデータをカテゴリーに分類するためパタンの種類を計算した。表 1 が示すように、全ての発話のうち約半分、つまり７５万発話は２度以上出現しない発話で、残りの約半分は繰り返し出現した発話である。その繰り返しパタンには 4 千種類以上あることが分かった。たとえば、パタン「うん」という発話は 4 万 8 千回現れた。パタン「あー」という発話は独立発話として、15 万回出現した。

＜＜表 1 ここ＞＞

表２では、500 回以上使われた発話パタンを出現頻度によりリストアップした。多くの場合、この文字列のみでは話者の発話意図は判断できない。たとえば、「うん」は肯定・否定・躊躇などのいくつかの意味解釈が可能で、そのときの言い方によって、話者意図が示される。このような発話パタンを以下「A-タイプ」と呼ぶ。繰り返し現れることのない発話パタンについては以下「I タイプ」と呼ぶ。「A」は Affect（情動情報）を、「I」は Information をそれぞれ指し、この分類により 2 種類の発話意図を区別することが可能になる。I-タイプ発話は主に言語的情報を持つが、A-タイプ発話は主に情動情報を示す [7]。A-タイプ発話の多く

は短発話であり、文字情報だけでの解釈は不可能である。表２が示すように同一くりかえしがよくみられる。たとえば、うん 48038 回、うんうん 1550 回、うんうんうん 1246 回、うんうんうんうん 599 回など...

＜＜表 2 ここ＞＞

A-type, I-type の区別と,「提供」「依頼」の発話行為を組み合わすと、4 種類の発話意図が考えられる。つまり、情報提供、情報依頼、アフェクト表示、アフェクト共有（依頼）の 4 種類であり、それぞれの組み合わせを表３で示す。これらの発話意図は以下、スピーチイベント（event） とも呼ぶ。

4　発話様式を説明する枠組み

これらの知見をまとめる意味で、対話発話様式を定義する新しい枠組みを提案する。「U = E | (S,O)」は「発話またはある表現形式（U: utterance）は、発話意図 E（event）の実用化であり、その基本条件は S と O に依存し、S（self）は自分の次元、O（other）は相手の次元である」ことを表す（図３）[8]。図は、簡略化しすぎではあるが、模式的に 2 段階ずつを示している。

自分の次元は「興味」と「気分」、相手の次元は、「聞き手との関係」と「環境状況」である。話者の気分がよく、興味のあるはなしで、親しい人と楽な環境で話をするという条件のとき、最も明るい話し方になる。もし相手が親しくない場合、或いは楽な環境でない場合、 話し方が極めて硬くなると考えられる。話し手の属性、話し手と聞き手の関係に 基づく発 話の分類認識（E）は指向性を持つと 考えられ、話し手の属性 （S） と聞き手と の関係 （O） の枠組みに基づいて I-タイプまたは A-タイプの情報をやり取りする機能を 持っている。

＜＜図 3 ここ＞＞

ここで、再び言葉と意味の関係に話を戻す。ある発話を言い換える、つまり別の文表現に 置き換えたり、翻訳した場合、その文形式は変わるが話者意図はそのまま残る。それに対 して、言い方を換えれば、言葉はそのまま残るが、意味や意図が大きく変わる。対話におけるこのような「言葉の遊び」は人間の特徴の一つである。つまり我々は言葉の選択、言い方の選択、特に上述のような非語彙的短発声をもって意見や意図を示すことができる。

Table 3: Basic utterance types for the *Event* category

	seeking	offering
I-type	interrogative	declarative
A-type	back-channel	exclamative

4.1　ラッパーとフィラー

本節では上述の A-タイプ、I-タイプの区分を拡張し、2 つの発話タイプが互いに混じり合い、同時に現れることが、自然相互的発話の特徴である”ill-formedness”を作り上げていることを示す。

上記で述べた枠組みで A タイプ発話の心的情報の示し方は説明できるが、I タイプについては充分ではない。その理由として、これまでの理論では、I タイプ発話は言語的情報だけしか示さないということになるが、実際は I タイプ発話も情動情報・話者の感情情報などを示す場合があることがあげられる。たとえば、「明日、九州に行く。」は文字列だけで充分意味が伝えられるため、I タイプ発話となる。もちろんその言い方によって話者の気分も分かる。しかし日常的な対話において、上記の文章は「あんな、明日な、九州に行くんよ」のようになる。このようにどちらの要素も兼ね備えているがゆえに、ある発話が A-タイプであるか I‐タイプであるかを判断することが困難である場合がある。したがって，あらゆる発話をこの 2 つのタイプに容易に分類することはできない。上記の例文「あした」は前後に装飾語を伴う。「あんな」と「な」は自然につながり、話者の心的情報が表出する。

また同様に文末の「んよ」も同じ機能を持つ。この装飾的と考えられる書き言葉には現れない話し言葉の特徴をここでラッパー（wrapper）と呼び、その他をフィラー（filler）と呼ぶ。以下に，典型的ともいえる日常会話における語及び句の現れかたを示す（図３）。

ここではフィラーを,「すき間を埋めるもの」或いは「談話における空白部分を補うもの」といった通常の解釈とは，意図的に異なった意味として用いる。本稿では無計画に産出されしばしば非言語的で ある音が，非常に高頻度に発話内に現れることから，これらの音がただ単に発話産出に必要な時間を 提供するだけでなく、声質と発話様式の特徴を捉えやすく、また話者の心的状態情報の指標としての役割をも果たしていると考えられる。

4.2　「A-タイプ」「I-タイプ」の自動識別

I-タイプ発話と A-タイプ発話の最も大きな違いは、その長さである。A-タイプの中には「おはよう」、「元気？」、「昨日の試合見た？」といった比較的長い発話が含まれる場合もあるが、これらがやはりあいさつ言葉であることに変わりはない。I-タイプ発話と定義されるためには、もっと多くの語が連なり、より長い発話を形成する必要がある。

本稿では，そのようなより長い発話に見られる非言語的断片のバラエティと出現頻度を調べ、それらをラッパーとフィラーに区別し更なる下位分類を試みた。言語学的知見からではなく，高頻度で現れるラッパーの辞書をつくるにあたり、”longest-common-substring” アルゴリズムを用いた。

このアルゴリズムによって、書き起こしされたコーパスにおける最も出現頻度の高い記号の連続が、発話開始点または発話終了点に現れていることが確認できる。トレーニングデータとして、ひらがな 20 文字から 40 文字の長さを含む発話を用いた。10 回を最低基準とし、それ以上繰り返し現れるラッパーに一致する文字列を、まず発話の左（開始点）から右（終了点）に向かって検索し、発話開始点に高頻度で出現する 899 個を抽出した。次に発話を右（終了点）から左（開始点）に向かって検索し、発話終了点に現れるラッパーを 957 個抽出した。

これらの「発話の両端」に現れるラッパーは、発話開始点で生起しようと、比較 的長い発話をラッパーとフィラーに分ける分節点で生起しようと、発話開始点及び終了点に現れるものはラッパーであり、発話内に現れるものはフィラーである。

図 4 は上記の 2 段回処理を行った結果を示している。太字で示した「語」は、非語彙的である典型的なラッパーである。(これらのラッパーが高頻度で現れていることは、恐らく日本語の文字を全く読むことができない人でもテキストを見れば理解できるだろう。)「#」で始まっている行は、それぞれの自然発話を標準的な日本語に書き直したものであり、これらはほとんどラッパーを含んでいない。自然会話における 1000 発話からは、 2337 個のラッパーが抽出され、各発話に対する平均生起回数は 2.34 回であった。書き起こしの形態素分析なしに 1 文字からなるラッパーを抽出することが困難であることを考慮すると、ラッパーの実際の生起度数はさらに高くなることに注目されたい。

＜＜図 4 ここ＞＞

5 まとめ

本稿では、一般に「くだけた」と思われている話し言葉の特徴を ”broken ”や ”ill-formed ”ではなく、ラッパーとフィラーという観点から捉えなおし、それらが対話において重要な役割を果たしていることを示した。ラッパーは、出現頻度が高いことと、多量の言語的内容を担うことがないため、聞き手にとって話し手が未知の人であっても、それが持つパラ言語情報は言語的情報と同じほどに受け取りやすいものとなる。書き言葉やよく計画された発話では無視すべきと思われる装飾的な A-タイプ発話が、逆に重要な役割をもつと考えられる。このことは、音声文法が書き言葉のそれとは異なるものであり、書き言葉の判断基準をそのまま用いることができないことを示している。音声言語のコミュニケーションは気持のやり取りを含む。

6 謝辞

本研究は科学技術振興機興 (JST) 及び情報通信研究機構 (NICT)、総務省戦略的情報通信研究開発推進制度 (SCOPE)、科学研究費補助金の援助によるもので、また激励と支援をくださった ATR の皆様にも感謝の意を表する。さらに、早川久美子、中川明子、田畑あきこ、田中三恵子のアドバイスを感謝する。

References

[1] 「表現豊かな声の秘密」『ヒューマン・インフォマティクス』2005. 監修長尾真. 工作舎. ｐｐ.65 － 84

[2] Campbell, N., & Erickson, D., 2004. "What do people hear? A study of the perception of non-verbal affective information in conversational speech", in *Jnl Phonetic Society of Japan.*

[3] 「声質ーパラ言語情報を持つ第四のパラメーター」『文法と音声ー IV』 音声文法研究会、くろしお出版 pp.25-34.

[4] Alku P., and Vilkman, E., 1996. "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering", *Speech Comm.*, vol.18, no.2, 131-138

[5] Alku, P., Backstrom, T., and Vilkman, E., 2002. "Normalized amplitude quotient for parametrization of the glottal flow", *J. Acoust. Soc. Am.*, vol.112, no.2, 701-710

[6] Campbell, N., 2004. "Accounting for Voice Quality Variation", Proc 2nd Intl Conf on Speech Prosody, Nara, Japan. 217-220

[7] Campbell, N., 2005. "Getting to the Heart of the Matter; Speech as the Expression of Affect", *Language Resources and Evaluation*, Volume 39, Issue 1, 111-120

[8] Campbell, N., 2004. "Extra-Semantic Protocols; Input Requirements for the Synthesis of Dialogue Speech" in *Affective Dialogue Systems, Lecture Notes in Artificial Intelligence*, vol. 3068 Eds Andre, E.; Dybkjaer, L.; Minker, W.; Heisterkamp, P., New York, Springer. 221-228

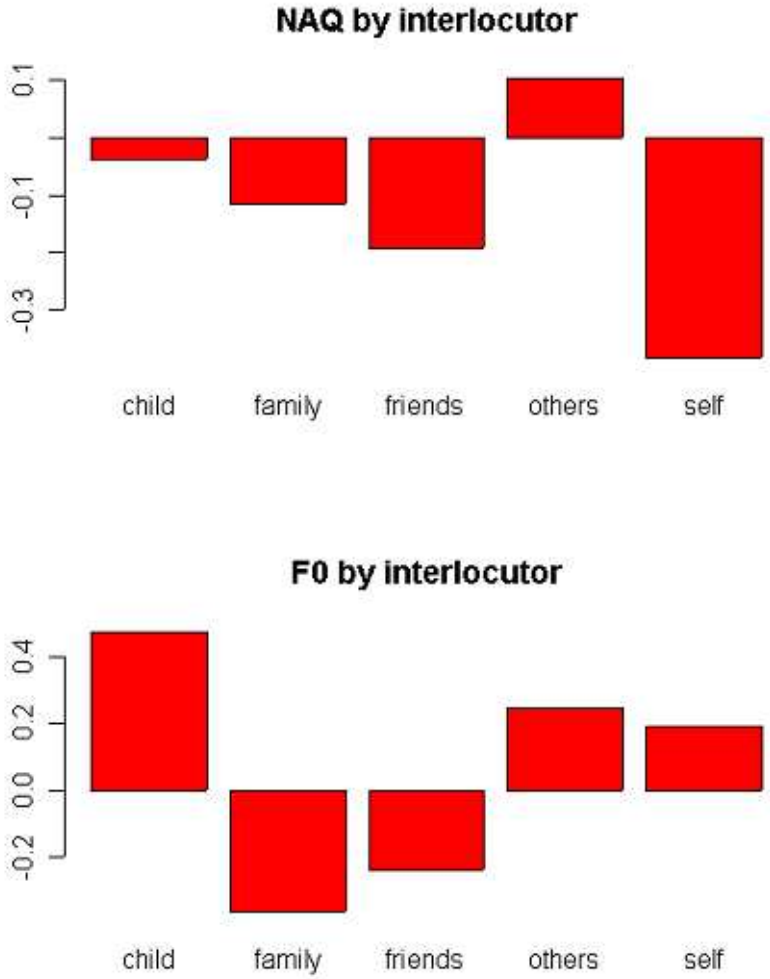

Figure 1: *Median values of NAQ and F_0 plotted for interlocutor. The data are (z-score) scaled, so values are in SD units. 0 represents the mean of the distribution*

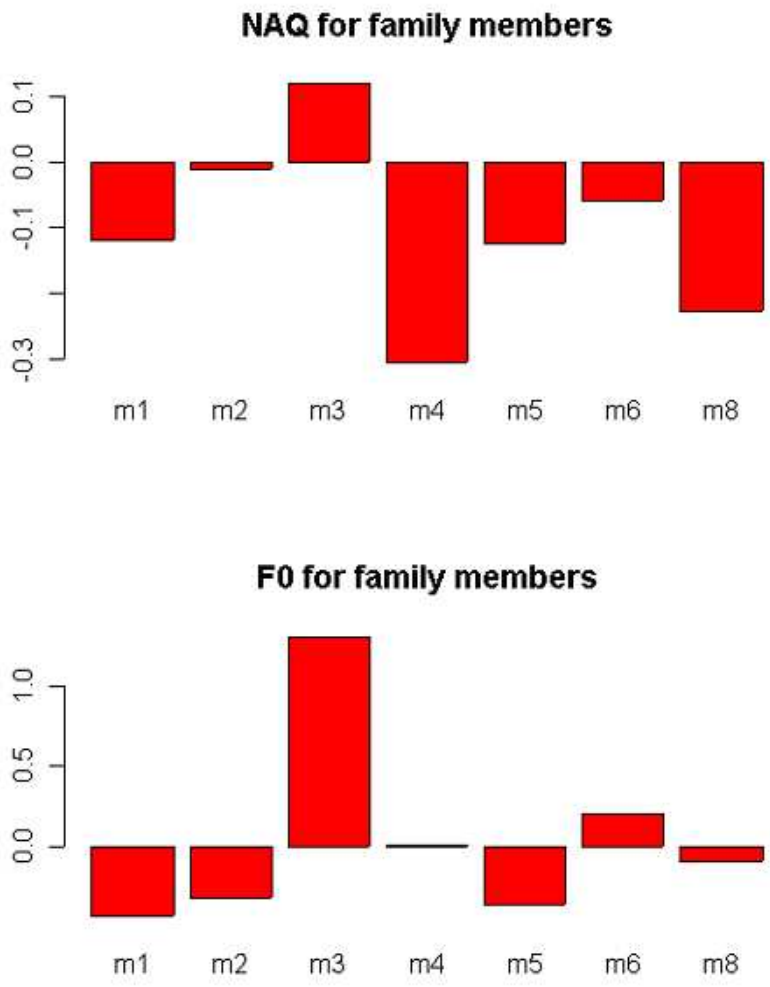

Figure 2: *Median values of NAQ and F_0 for family members. m1: mother, m2: father, m3: daughter, m4: husband, m5: older sister, m6: sister's son, m8: aunt*

あ、もしもし、あのちょっとけいやくのないようへんこうしていただきいんですけど
(もしもし。契約の内容を変更していただきたいのですが)
しらんゆう**ねんな**、ひつこい**ねんもう**ぴかぴかぴかぴかひかってるからきになってさあ
(知らないと言っているのに、しつこいぴかぴか光っているので気になって)
たべれん**ねん**で、たべれん**ねんけど**きもちわるいし、まだまだしんどいし、**みたいな**
(多分食べられます。食べられるのですが、気持ちが悪いしまだしんどい、と言った感じで)
だかほんまあんまたたかんでいいらしい**ねんけど**、**ま**、ちょっとほこりおとすていど
(だから本当に、あまり叩かなくて良いらしいです。少し埃を落とす程度で)
うんうんうん、でもさ、どうせさ、いろいろあつめんねやったら、これをしってたら
(どうせ色々集めるのなら、これを知っていれば)
うん、**そら**、こまま、こおりやまのほうちょっとまっすぐいったところ**やねんけどな**
(はい。このまま郡山の方へまっすぐ言ったところなのですが)
まあはんなどうろあるから**なあ**、やっぱりつうこうりょうすくないかもしれん**よなあ**
(まあ、阪奈道路があるからやっぱり交通量は少ないかもしれませんね)
それもかんがえようよな、**なんか**ほんまにきんてつでぜんぶすんねやったらいいけど
(それも考えようですよ。本当に近鉄で全部するのならいいけれど)
あるくのいたい、**む、なんか**どっちかはんぶんがすごいしびれてあるかれへん**ねん**て
(歩くのが痛い。どちらか半分がとても痺れて歩けないのです)
なんかめんどくさいな、おかしつねにかっとかなあかん**やんとか**おもっとってんけど
(何か面倒くさいなあ。お菓子は常に買っておかなければならないと思っていたのですが)
なんかさあ、あのかたちがちゃんとなってへんからはきにくいすりっぱってある**やん**
(形がちゃんとしていないために歩きにくいスリッパがあるじゃないですか。)

Figure 3: Sample utterances of Japanese conversational speech, selected at random from those having a length of between 20 and 40 mora in the corpus. Each utterance is followed by its equivalent transliteration in standard Japanese for comparison. Bold font shows the automatically-detected 'wrappers' in these utterances

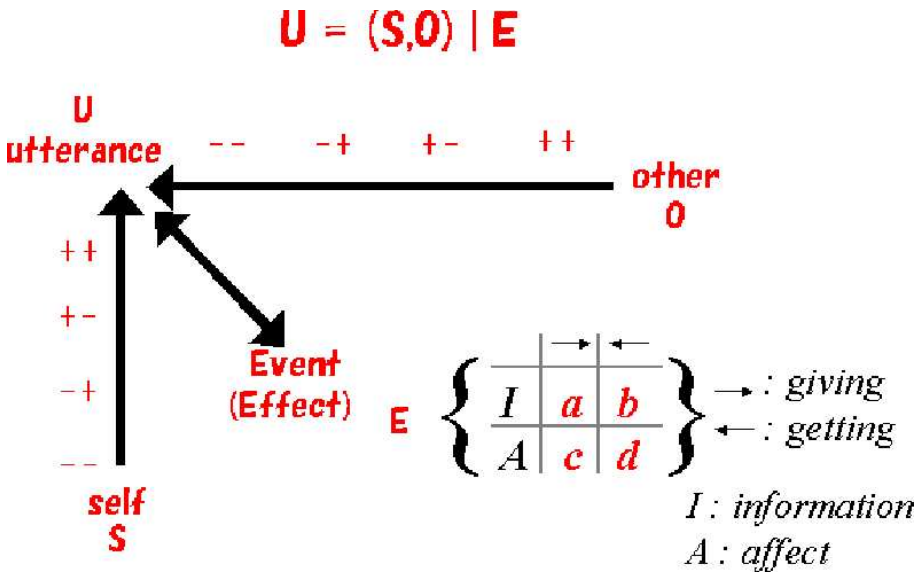

Figure 4: *A 3-dimensional framework for categorising speaking-style and expressiveness; an utterance is realised within the constraints of speaker (self) and interlocutor (other) according to the discourse intention (event). Voice quality, speech-rate, prosodic range, and voice power will vary accordingly*