加齢等による発話への影響の音声情報を用いた評価

松浦 博[†1]　鈴木大虎[†1]　井本智明[†1]　和田淳一郎[†2]　秀島雅之[†3]

概要：加齢等による咽喉の衰えを発話を評価することによって早期に検知し、自覚することによって対策すれば健康維持につながると考えられる。本報告では声帯の振動を反映する基本周波数 F0 に加えて，独自に開発した音声セグメントラベルを用いて求めた様々な発話に関するパラメータを抽出した。パラメータとしては，F0 変化幅，母音定常部比率，有声音部 F0 分散，音声区間長，無音区間長、妥当ラベル比率，かすれラベル，ラベル安定性を用いて評価した．本方法によって，発話の正確性，かすれ声や濁った声，抑揚の少なさ，たどたどしさ、安定性等の発話の特徴を評価することが可能であることが分かった．

キーワード：発話評価，加齢，高齢化，音声セグメントラベル，基本周波数

Evaluation of the influence of aging on utterance by using speech information

HIROSHI MATSUURA[†1]　DAIGO SUZUKI[†1]　TOMOAKI IMOTO[†1]
JUNICHIRO WADA[†2]　MASAYUKI HIDESHIMA[†2]

1. はじめに

咽喉の機能にはエネルギー摂取のために食べ物を飲み込む「嚥下」，酸素を肺に取り入れ二酸化炭素を排出する「呼吸」，他者とコミュニケーションを担う「発話」がある．いずれも人間の生命維持や社会生活にとって，極めて重要な役割を果たしている．人間が多くの動物と異なる点として，直立歩行，道具の使用などが良くあげられるが，真に人間だけと言えるのは複雑な言語情報を伴う音声である．脳の中での複雑な思考は言語を用いるからこそ可能になるのであって，発話は人間の本質であるとさえ言える．

これらのことからも，咽喉が人間の生命維持に，いかに重要な役割をしているかが分かる．長寿を全うする上においても，健全な発話を維持することは極めて重要である．人間が音声を使えることと裏腹に，食物を消化する食道と呼吸のための気道が咽喉において交差するため，唾液を肺に誤って送り込む誤嚥という危険が喉の衰えた高齢者に，もたらされる．また，誤嚥は嚥下の際にだけ起こるのではないため，食事中にだけに注意すれば良いというわけでもない．加齢等による咽喉の衰えはゆっくりと進行することが多いため，特に，本人はなかなか気づかないという問題もある．発話を評価することによって早期に咽喉の衰えを検知できれば，可逆的に対応可能と考えられている[1].

†1 静岡県立大学　経営情報学部
University of Shizuoka, School of Management and Information
†2 東京医科歯科大学　大学院医歯学総合研究科
Tokyo Medical and Dental University，Graduate School of Medical and Dental Sciences
†3 東京東京医科歯科大学 歯学部附属病院
Tokyo Medical and Dental University，University Hospital of Dentistry

嚥下機能を強化するための咽喉の運動と，その評価に音声を用いることはある程度行われているが，近年発展する音声技術を発話評価に取り入れた試みはほとんど行われていない．ここでは，声帯の振動を反映する基本周波数（F0）に加えて，独自に開発した音声セグメントラベル[2]（以下，ラベルとも表記）を活用して求めた様々な発話に関するパラメータによって，発話の状況を手軽に推定・評価する試みについて報告する．

2. 発話分析の方法

ここで用いた発話分析評価システム（以下，本システムとも記述）の基本的な処理の流れを図 1 に示す．本システムは音声セグメント，音声パワー，基本周波数 F0 を計算する特徴抽出部と発話音声を評価するための，F0 変化幅，母音定常部比率，有声音部 F0 分散，音声区間長，無音区間長，妥当ラベル比率．かすれラベル，ラベル安定性などのパラメータを求めるパラメータ抽出部からなる．

2.1 音声セグメント抽出

学習者が発話した音声はサンプリング周波数 22.05kHz，量子化ビット数 16 で量子化した後，フレーム長 23.2ms（ハミング窓），フレームシフト 8ms で 512 点の FFT（高速フーリエ変換）分析される．その後，32 チャンネルの BPF（バンドパスフィルタ）群により周波数パラメータを抽出する．この周波数パラメータの 6 フレーム分である時間－周波数パターン[2]を音声特徴パターンとして，音素あるいは音素間の遷移等を示す音声セグメントの標準パターンとマッチングする．なお，標準パターンはあらかじめ男女各 400 名程度の大量の音声データから求めている．

マッチングの結果，最大類似度を示す音声セグメントのラベルを図 2 の最下段のように 8ms 毎に時系列で表示している．この音声セグメントの標準パターンは男声・女声用に分けて用意するなど細かな区別をしているため，本来は 690 種と多いものの，これを表 1 に示すような 213 種類の音声セグメントラベルに統合している[2]．この 213 種類を改めて音声セグメントラベルと呼び，音声特徴を反映したアルファベット 2 文字で表記する．これよって，一目で理解できるようにするとともに，ラベルの縦方向の 2 文字でフレームごとに表示することを可能としている．

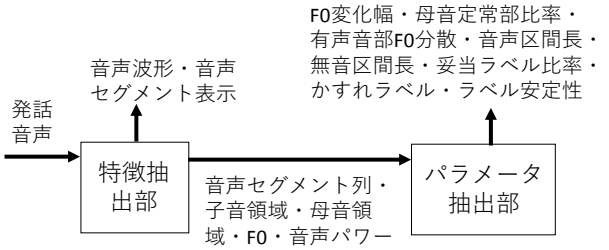

図 1　発語分析システムの構成

表 1　音声セグメントの一覧

セグメント 母音性	AA	EE	II	OO	UU	XX	AE	AI	AO	AU	AW	AX	AY	EA	EI	EO
	EU	EW	EX	EY	IA	IE	IO	IU	IW	IX	I4	OA	OE	OI	OU	OW
	OX	OY	UA	UE	UI	UO	UX	UY	U4	XA	XE	XI	XO	XU	XW	XY
子音性セグメント	CC	FF	HH	##	SS	$$	BB	DD	GG	ZZ	RR	I4	MM	NN		
	BA	BE	BI	BO	BU	BY	CI	CU	CY	DA	DE	DI	DO	DU	FA	FE
	FI	FO	GA	GE	GI	GO	GU	GY	HA	HE	HI	HO	HU	HY	JI	JY
	KA	KE	KI	KO	KU	KY	MA	ME	MI	MO	MU	MY	NA	NE	NI	NO
	NU	NY	PA	PE	PI	PO	PU	PY	QA	QE	QI	QO	QU	QW	QY	RA
	RE	RI	RO	RU	RY	SA	SE	SO	SU	SY	$I	TA	TE	TI	TO	WA
	WE	WI	WO	YA	YE	YO	YU	ZA	ZE	ZO	ZU					
その他のセグメント	QQ	BZ	AB	AD	AG	AH	AM	AN	AQ	AR	AS	AZ	EB	ED	EG	EH
	EM	EN	EQ	ER	ES	EZ	IB	ID	IF	IG	IH	IM	IN	IQ	IR	IS
	IZ	OB	OD	OG	OH	OM	ON	OQ	OR	OS	OZ	UB	UD	UG	UH	UM
	UN	UQ	UR	US	UZ	XB	XD	XG	XH	XQ	XR	XS	XZ			

注）　I4は母音領域では母音性セグメント,子音領域前では子音性セグメント

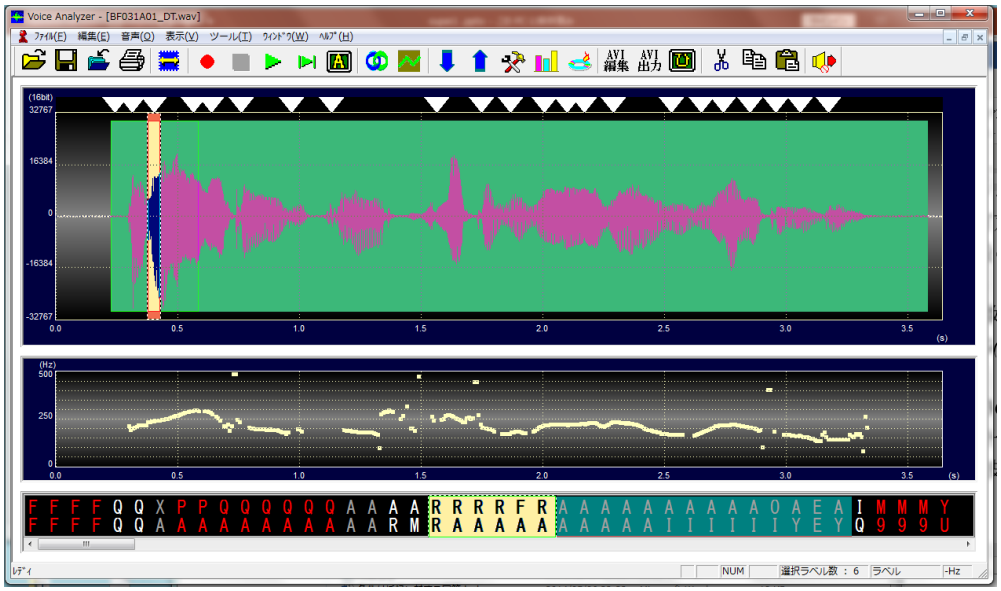

図 2　特徴抽出部の画面（発話内容：あらゆる現実をすべて自分の方へ捻じ曲げたのだ）

2.2　子音領域と母音領域

音声セグメントラベル列から表 1 に示す子音性セグメントが 3 フレーム(24ms に相当)以上で連続する領域を子音領域として抽出する．3 フレームとしたのは多くの実データから導き出した閾値であるが，子音の継続時間は短いもので 20ms～50ms とされている[3]こととも矛盾しない．子音領域を推定したのち，後続する母音領域を母音性セグメントが 3 フレーム以上の連続する領域として抽出する．母音は短い子音よりは一般的に長いものの，発話が速い場合や，発話が乱れたために母音性セグメント以外が挿入され，母音領域が分かれて短くなる場合もあるため，閾値を 3（フレーム）とした．

図 2 に示すように，上段に波形データが表示され，中段に基本周波数 F0 を表示している．下段に音声セグメントラベルを 8ms ごとに表示している．波形データの上にある逆三角マーク▽をクリックすると，それぞれの子音領域を指定できる．指定された子音領域に対応する音声波形ウィンドウ，音声セグメントウィンドウがハイライトされるため，確認しやすい．図 2 に示す例では音声セグメントウィンドウの/ra/の子音領域が黄色でハイライトされ，かつ中心

に表示されている．それに続く母音領域/a/が，緑色で表示されている．なお，音声セグメントラベル列の表示されていない部分は一番下のスクロールバーを移動すれば確認できる．

表 1 の母音性セグメント，子音性セグメントの分類はそれぞれ母音領域，子音領域を後述の方法で求めるために設定している．母音性セグメントには，5 母音に対応する AA, EE, II, OO，UU のほか，撥音 XX，母音・撥音・半母音間の遷移部を示す AE, AI など 48 個である．なお，I4, U4 はそれぞれ II, UU の無声化を示している．

子音性セグメントに関して，/chi/および/cu/, /fu/, /hi/, /shi/の子音部分をそれぞれ CC, FF, ##, $$で示す．また，ハ行，サ行，バ行，ダ行，ガ行，ザ行，ラ行，マ行，ナ行のそれぞれの子音部分を，HH, SS, BB, DD, GG, ZZ, RR, MM, NN で示す．ここで，子音性セグメントには子音から母音へ遷移する音声セグメント BA, BE などを含めているため，子音領域には純粋な子音部分だけでなく子音から母音へ遷移する部分が含まれることになる．さらに，母音性および子音性セグメント以外をその他のセグメントと分類した．

2.3 基本周波数 F0 の導出

F0 は無声音では基本的には存在しない．2 文字からなる各音声セグメントラベルの 2 文字目が母音・有声子音等であり有声音であることを示すフレームに限って，図 2 の中段のように表示した．*F0* の導出の概略を次に示す．

(1) 音声データ *p(k)*から自己相関関数 *G(L)*を次式によって計算する．自己相関係数 *G(L)*とは *p(k)*と *L* 個ずれた *p(k +L)*との積（相関）を，ラグ *L*=42（*F0* で 524Hz に相当）から *L*＝367（*F0* で 60Hz に相当）まで変化させてすべて計算する．なお，*F0* の推定範囲はサンプリング周波数 22025Hz／*L* から計算することができる．

$$G(L)=\sum_{k=0}^{N-L-1} p(k)\,p(k+L)$$

ここで音声データから逐次，切り出して分析する範囲となるフレーム長は *N*=512 である．

(2) *L* の関数 *G(L)*が最大となる *L*= L_{max} で音声データに基本的な周期性が認められるため，これが声帯振動を反映していると推定される．

(3) L_{max} から *F0* を次式によって求める．

$$F0 = 22025 / L_{max}$$

なお，音声の大きさを示すパワーは，(1)で述べた式で *L*=0 とした式と同様の次式である．

$$G(0)=\sum_{k=0}^{N-1} p(k)\,p(k)$$

2.4 実験データ

実験データには，20 代～50 代の話者が発話した日本音響学会新聞記事読み上げ記事コーパス（以下， JNAS）と，60 代以上の高齢者が発話した新聞記事読み上げ高齢者音声コーパス（以下，S-JNAS）を用いた．ここでは音素バランス文の最初の「あらゆる現実をすべて自分の方へ捻じ曲げたのだ」と「冬が長くてつらければ，それだけ喜びも大きいのだ」を被験文とした．前者の被験文に関しては JNAS は男性 15 名と女性 16 名，S-JNAS は男性 25 名と女性 28 名である．後者の被験文に関しては JNAS は男性 14 名と女性 16 名，S-JNAS は男性 29 名と女性 29 名である．あらかじめ各データを聴取し，聴取者 1 名が明瞭さを 3 段階で評定した．発話からは様々な特徴を感じられるため，明瞭さという一つの指標で判断するのは難しいが，1 次評価として実施した．

2.5 発話評価用パラメータ

音声セグメントラベル，基本周波数 F0，音声パワー等を用いて抽出したいくつかの発話評価用のパラメータの抽出方法を次に述べる．

(1) **F0 変化幅**

被験文中の無声音・無音の部分を除いたフレームの F0 から最大値と最小値を求める．それらの差を F0 変化幅とし，文全体の抑揚の大きさの指標とする．

(2) **母音定常部比率**

ラベルが AA, II, UU, EE, OO となるフレームは安定した母音定常部とみなせる．この部分のフレーム数を(4)の音声区間長で割って比率を求める．

(3) **有声音部 F0 分散**

有声音を示すラベルが出力されるフレームの F0 の分散を求める．この分散は隣り合ったフレームでの F0 の差を 2 乗して足し合わせた値を，対象となったフレーム数で割って平均化して求めている．

(4) **音声区間長**

発話全体の音声パワーの最大値を求める．最大値の 0.1%より大きい音声パワーでかつラベルが QQ ではないフレームを発話フレーム，それ以外を無音フレームとする．発話フレームから発話の始端と終端を求め，終端から始端までのフレーム数を音声区間長とする．

(5) **無音区間長**

音声区間長から発話フレーム数を引くことによって，音声区間中で発話されていないフレーム数を無音区間長として求める．

(6) **妥当ラベル比率**

日本語のモーラは子音領域，母音領域と次の子音領域までの境界領域に分解できる[2]．各領域で出現すべき適切なラベルをあらかじめ設定する．例えば，「あらゆる」のモーラ「ら」の子音領域では RR, RA であり，「ら」の母音領域および境界領域では AA, AY が適切なラベルとする．自動抽出されたモーラでは付加・脱落誤りが起こりうるので，DP マッチングを用いて，各領域で適切とされるラベルの出現比率を累積し，その最大値を妥当ラベル比率とする．

(7) **かすれラベル**

ハ行の発話以外の部分でラベル HH, HA, HI, HU, HE, HO, ##, FF, AH, IH, UH, EH, OH, XH, HY のいずれかが出力されたフレーム数をカウントする．かすれた発話でこれらのラベルが出力されることがあることを利用している．「あらゆる現実をすべて自分の方へ捻じ曲げたのだ」では，「ほ」の発話以外で，「冬が長くてつらければ，それだけ喜びも大きいのだ」の冒頭の「ふ」の発話以外で，カウントした．

(8) **ラベル安定性**

アルファベット 2 文字から構成される各ラベルが隣接フレームで一致している場合は 2 とし，1 文字が一致している場合は 1 として累積し，フレーム当たりの平均値を求める．なお，ゆっくりと発話するとラベルは安定しやすく有利となるため，さらに音声区間長で正規化した．

3. 分析結果と考察

各パラメータについて 5 点を満点として正規化し，図 3 から図 12 にレーダーチャートで示した．外側に広がっているほど良い傾向であると考えられようにした．値が小さい方が良い傾向と考えられる有声音部 F0 分散，音声区間長，無音区間長，かすれラベルは大小を逆転させている．

表 2 はすべてのデータを用いて，明瞭さとの相関係数を求めたものである．無音区間長，妥当ラベル比率，ラベル安定性の相関係数が比較的に高い．かすれラベルは「あらゆる現実を・・」について相関係数が比較的に高い．

図 3 から図 7 は被験文「あらゆる現実を・・」についてのデータである．図 3 と図 4 はそれぞれ JNAS，S-JNAS で明瞭さが高い 3 と聴取された発話データから得られた各パラメータの値である．図 3 から母音定常部比率は 5 から 0 までと広範囲に分布していることからも，明瞭さへの影響は少ないのではないかと考えられた．また，図 4 の S-JNAS でも同様の傾向である．

表 2　明瞭さと各パラメータとの相関係数

パラメータ	F0変化幅	母音定常部比率	有声音部F0分散	音声区間長	無音区間長	妥当ラベル比率	かすれラベル	ラベル安定性
あらゆる現実を・・	0.40	0.08	0.26	0.46	0.32	0.49	0.53	0.52
冬が長くて・・	0.21	0.04	0.08	0.47	0.25	0.51	0.27	0.47

図 5 は JNAS で明瞭さが低い 1 と聴取された発話データである．BM117 はかすれがある印象であが，パラメータとしては，一般的な S-JNAS のデータと比較すると悪くはないが，ラベル安定性や有声音部 F0 分散が低い．BM119 はかすれと「自分の」の部分が不正確に感じられ，特に「ぶ」が「ぷ」に近い発話になっている．いずれも発話速度は速いので，たどたどしさは感じない．

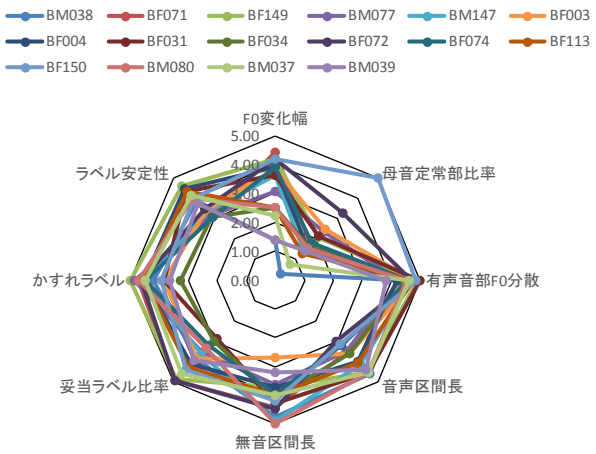

図 3　明瞭さ 3 の JNAS 被験者のパラメータ「あらゆる現実を・・」

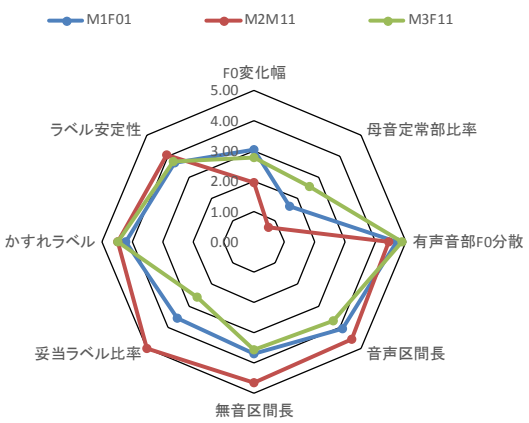

図 4　明瞭さ 3 の S-JNAS 被験者のパラメータ「あらゆる現実を・・」

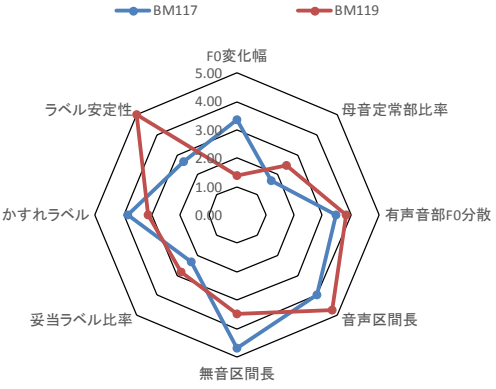

図 5　明瞭さ 1 の JNAS 被験者のパラメータ「あらゆる現実を・・」

図 6 に示す明瞭さ 1 となった S-JNAS の女性データでは，かすれた印象であり，かすれラベルにこの傾向が見られる．一方，明瞭さ 1 となった図 7 の S-JNAS の男性データでは濁った印象で，有声音部 F0 分散にこの傾向が見られる．また，一般的に音声区間長が低い値になるほど，ゆっくりとした発話を示し，そのため年齢を感じることがある．一方，ゆっくりとした発話では母音部が長くなるが，丁寧に発話された場合に，母音部で妥当なラベルを稼ぐことになるため，妥当ラベル比率が大きくなる状況も見られた．

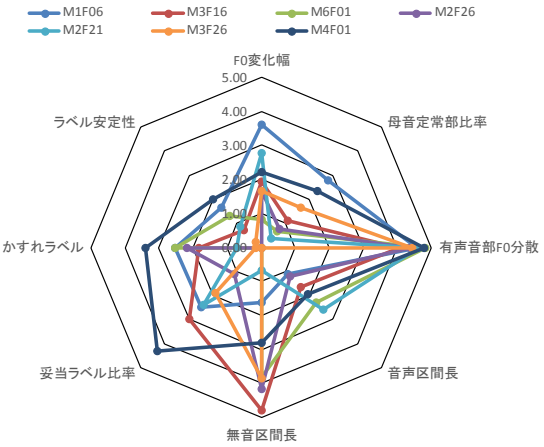

図 6　明瞭さ 1 の S-JNAS 女性被験者のパラメータ「あらゆる現実を・・」

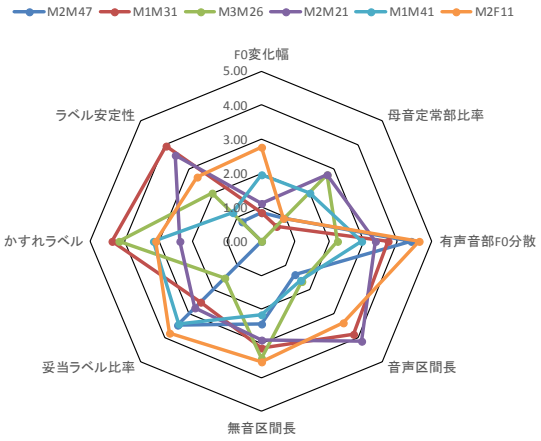

図 7　明瞭さ 1 の S-JNAS 男性被験者のパラメータ「あらゆる現実を・・」

図 8 から図 12 は被験文「冬が長くて・・」についてのデータである．図 8 と図 9 はそれぞれ JNAS，S-JNAS で，明瞭さが高い 3 と聴取された発話データである．図 8 の BF027 は妥当ラベル比率は低いものの、音声区間長は短く，かすれラベルも少なく活発な印象が捉えられている．図 9 は妥当ラベル比率がいずれも高めである．BM127 は有声音部 F0 分散、ラベル安定性、F0 変化幅が低めで濁った印象であるが，かすれラベルや妥当ラベル比率は高めで，総合的には明瞭さ 3 となっている．

図 10 は JNAS で明瞭さが 1 と聴取された発話データである．BF104 は特に冒頭があいまいな発話で妥当ラベルが低くなったと考えられる．

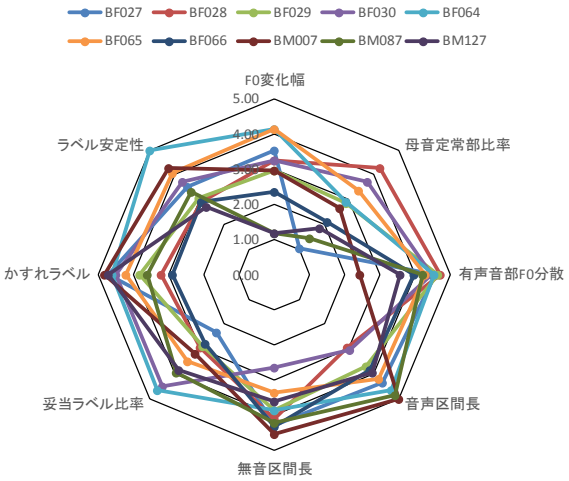

図 8　明瞭さ 3 の JNAS 被験者のパラメータ「冬が長くて・・」

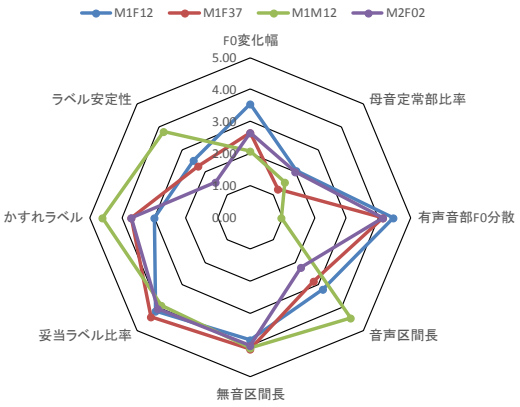

図 9　明瞭さ 3 の S-JNAS 被験者のパラメータ「冬が長くて・・」

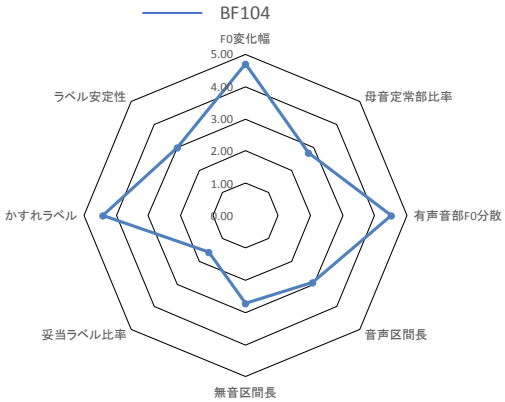

図 10　明瞭さ 1 の JNAS 被験者のパラメータ「冬が長くて・・」

図 11 は S-JNAS 女性被験者で明瞭さが 1 と聴取された発話データである．例えば，M1F22 は息継ぎが極めて長いため，無音区間長の値が 0 で，有声音部 F0 分散も低く不安定な発話という印象になったと考えられる．

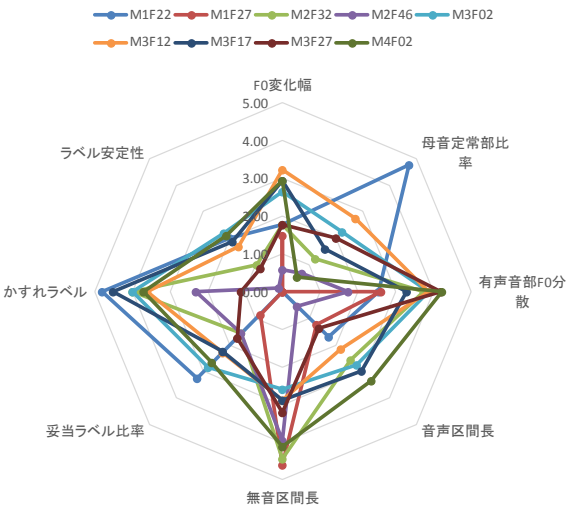

図 11　明瞭さ 1 の S-JNAS 女性被験者のパラメータ
「冬が長くて・・」

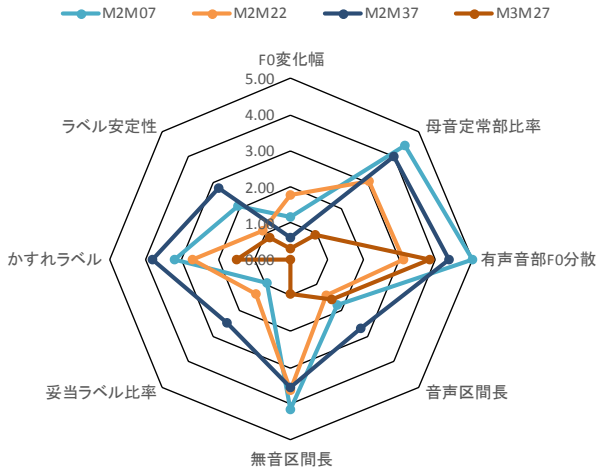

図 12　明瞭さ 1 の S-JNAS 男性被験者のパラメータ
「冬が長くて・・」

4.　おわりに

様々なパラメータによって加齢による発話や咽喉の衰えに関わるかすれ声や濁った声，抑揚の少なさ，たどたどしさ等の発話の特徴を表現することが可能であることが分かった. また, 発話がゆっくりであれば明瞭さは増す一方で，たどたどしい印象になるなど，聴取評価によって，一つの指標で明瞭さを付与するのは容易ではないことも分かった．今後は，被験文を増やしてパラメータをさらに検討し，深層学習による発話自動評価も実施する予定である．また，聴取者による聴取評価の精度の改善も検討したい．

謝辞　本研究の一部は科研費(16K00484)の助成を受けて実施した．

参考文献

[1]　https://www.jda.or.jp/enlighten ment /oral/about.html.
[2]　松浦博，桃崎浩平，正井康之，秀島雅之，犬飼周佑，佐藤雅之，安藤智宏，大山喬史，” チェアサイドで使用可能な発語評価のための音声認識装置の開発，” 情処学論, vol.46, no.5, pp.1165-1175, 2005.
[3]　板橋秀一，” 音声工学，” 森北出版, p.33, p.36, p.46, p.104, p.159, 2005.