音声認識技術を用いた端末における音声検出部の低消費電力化

EMA, Kentaro / 江馬, 健太郎

(出版者 / Publisher)
法政大学大学院理工学研究科
(雑誌名 / Journal or Publication Title)
法政大学大学院紀要. 理工学・工学研究科編
(巻 / Volume)
61
(開始ページ / Start Page)
1
(終了ページ / End Page)
5
(発行年 / Year)
2020-03-24
(URL)
https://doi.org/10.15002/00022817

音声認識技術を用いた端末における音声検出部の低消費電力化

LOW POWER CONSUMPTION OF SPEECH DETECTION UNIT IN TERMINALS USING VOICE RECOGNITION TECHNOLOGY

江馬健太郎
Kentaro EMA
指導教員　安田彰

法政大学大学院理工学研究科電気電子工学専攻修士課程

In recent years, devices using voice as a user interface have been increasing. These devices consume a lot of power and are not a problem when using an external power supply, but this is a serious problem with battery-powered devices such as IoT devices. So we propose an analog frontend circuit combining a speech detection and position detection to improve power consumption, which reduces active operation period of analog circuits. This system consists of a zero-crossing method or LPC analysis and a position detection. The analysis shows that the relation among the power consumption, detection time, and detection accuracy.

***Key Words** : Speech recognition, Speech detection, Position detection*

１．はじめに

近年，音声認識技術の発展に伴い，音声信号をユーザーインターフェースとして動作する機器が増えており，今後も増え続けていくと考えられる．このようなシステムでは複数のマイクから常に音声信号を取得するために，複数個の高精度な Analog-to-digital converter (ADC)を常時動作させ続ける必要がある．また，信号取得後にも複雑な音声処理やパターン識別などを行う必要がある [1][2]. これは消費電力の増大に繋がり，外部電源を用いているデバイスならば問題にならないが，IoT デバイスなどのバッテリー駆動のデバイスにおいては重大な問題となる．そのため音声認識デバイスの低消費電力化を目的とした，マルチステップ型の音声検知システムの検討は，有用な研究課題と考えられる．

本稿では，このシステムの中でも音声検出部に使用する音声検出法としてゼロクロス法および LPC 分析法について MATLAB を用いてシミュレーションを行った．

第 2 章では現在検討しているマルチステップ型の音声検知システムについて述べる．第 3 章では検討する音声検知技術であるゼロクロス法および LPC 分析についての解説を行う．第 4 章では検討中のマルチステップ型の音声検知システムに第 3 章で述べた音声検知技術を適応する場合においての懸念点を述べる．第 5 章では第 4 章で述べた懸念点についてゼロクロス法および LPC 分析で行ったシミュレーション内容および結果を示す．第 6 章では本稿のまとめおよび今後の課題について整理，考察を行う．

２．マルチステップ型の音声検知システム

現在，私たちは音声認識デバイスの低消費電力化に対して，マルチステップ型の音声検知システムを導入し，音声認識の稼働時間を減らすことにより消費電力を低減させるシステムの検討を行っている．次の図 1 に検討中のシステム構成を示す．

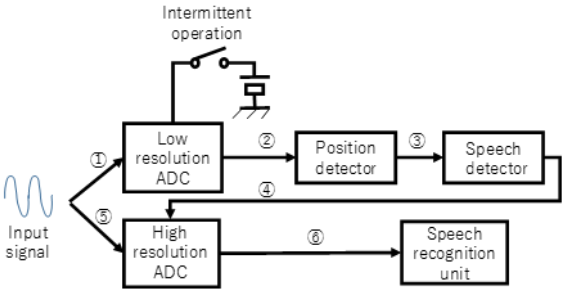

図 1　検討中の音声検知システム

このシステムは図 1 に示したように，まず物体の位置検知を行い，何かを検知できた場合はその後に音声検知を行う．これにより音声が検知できた場合のみ高精度な ADC を立ち上げ音声認識を行う．また音声認識を立ち上げた後に，音声を一定時間認識しなかった場合には位置検知を行う．これにより何も検知できなかった場合は位

置検知のみを行う最初の状態に戻るという構成を想定している．

位置検知のみでは人以外のものを検知した場合においても音声認識を立ち上げてしまう問題点があった．また音声検知のみでは人が近くにいない場合でもテレビやラジオなどで音声が流れている場合においても音声認識を立ち上げてしまう問題点があった．これにより，どちらか 1 つでは音声認識の稼働時間を大きく削減することは見込めないと考えられる．しかし，この 2 つを組み合わせることで互いの問題点を補うことができるため，音声認識の稼働時間を削減し必要最小限に近づけることが見込める．

３．検討する音声検知技術

（１）ゼロクロス法

ゼロクロス法とは，次の図 2 のように入力信号が 0 を通過した回数（ゼロクロス回数）を用いて入力信号の周波数を算出する方法である．これを用いて音声を検知する場合，算出した周波数について人の出しうる周波数帯と比較を行う．更に周波数のみでは雑音でも音声であると判断してしまう可能性があるため，入力信号のパワー平均を求め，これについても人の声が取りうる値と比較を行う [3]. この 2 つの判断基準により入力信号が音声であるかを判定する分析法である．

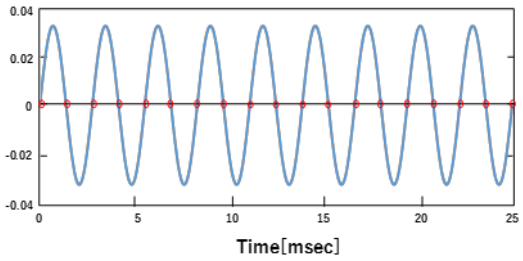

図 2　ゼロクロス検出

主にゼロクロス回数の算出には入力信号の符号の変化が用いられ，それを次の式(1)へ代入し入力周波数を算出する．音声であると判断する帯域の基準に関しては経験的なものであるため，今回のシステムでは 250 Hz から 1200 Hz を暫定的な基準として設ける．

$$\text{周波数[Hz]} = \frac{\text{ゼロクロス回数}}{2} \times \frac{1}{\text{フレーム長[sec]}} \qquad (1)$$

式(1)においてフレーム長とは処理する入力信号の長さのことである．また，パワー平均の算出には次の式(2)に示す *RMS*（二乗平均平方根）が用いられる．今回のシステムでは，*RMS* から *SPL* を算出し，これについて 58 dB SPL から 72 dB SPL を暫定的な基準として設ける．

$$RMS(n) = \left[\frac{1}{n}\sum_{m=0}^{n-1} S(m)^2\right]^{\frac{1}{2}} \qquad (2)$$

この式(2)において，n はサンプル点数であり，S は入力信号である．この 2 つから算出した値がそれぞれ音声の取りうる周波数およびパワー平均であった場合に入力信号を音声として判定する．

この手法のメリットとしては処理の軽さが挙げられる．これは LPC 分析などに比べてフーリエ変換などの重い処理を行わずに判定が可能なためである．また，デメリットとしては検出精度の低さが挙げられる．これはゼロクロス法における音声の判定基準が入力信号に音声の特性があるかを見ているわけではないためである．これによりノイズに対して音声と判定してしまうことや，音声である場合においても *RMS* が低い場合は検出できないことがあるため検出精度が低くなる．

（２）LPC 分析

LPC 分析とは，声道を音響管に見立て音声信号について次に示す式(3)が成り立つときの線形予測誤差を最小にするように線形予測係数を求めることにより，声道特性のスペクトル包絡を求め音声検知を行う分析法である [4].

$$x(n) - \{\alpha_1 x(n-1) + \cdots + \alpha_p x(n-p)\} = e(n) \qquad (3)$$

この式(3)において，$x(n)$は音声信号，$e(n)$は線形予測誤差，α_pは線形予測係数である．これにより求めたスペクトル包絡において周波数が低い方から，1 つ目と 2 つ目のピーク（フォルマント）の周波数を後述するフォルマントチャートと照会することにより，入力信号に母音の性質を含んでいるかを判定することで音声であるかを判定する．

この手法のメリットとしては，高速解法（Durbin の再帰的解法）を用いることで単純な操作でスペクトル包絡が抽出可能であること．声道特性から入力信号に音声の性質があるかの判定を行うため判定精度が高いことが挙げられる．また，デメリットとしては線形予測係数を求める過程で行列式を解く必要があるため処理が比較的重くなることが挙げられる．

（３）フォルマントチャート

（２）で求めたスペクトル包絡のピークについて，周波数の低いほうから第 1 フォルマント周波数，第 2 フォルマント周波数，…と呼び，第 1 および第 2 フォルマント周波数を次の図 3 に示すフォルマントチャートと照会することにより音声であるかを判断する．

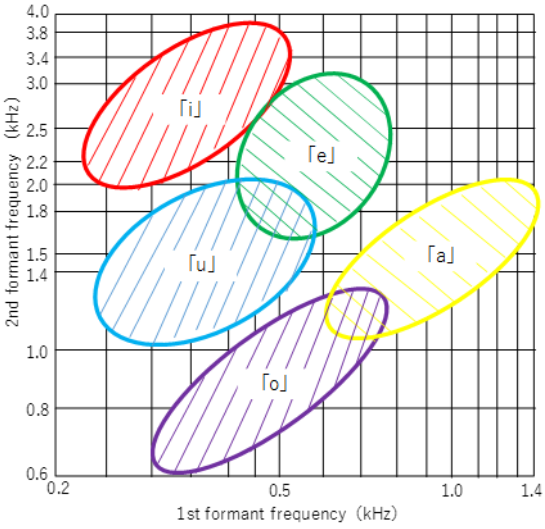

図 3　フォルマントチャート

図 3 に示したように，フォルマントチャートとは横軸に第 1 フォルマント周波数，縦軸に第 2 フォルマント周波数を取った場合において，日本語の母音が発声されたときに周波数がどの範囲を取るのかを示しているグラフである．このように各母音の取りうる周波数範囲には大きな幅があり，LPC 分析ではこの周波数範囲をどう設定するかも認識精度を変える要因の一つである．

4．提案システムにおける懸念点および検討内容

ここでは音声認識を用いた端末にこのシステムを適応させることを考える上での懸念点について述べる．

システム全体としては，位置検知と音声検知を順次行うため音声認識を立ち上げるまでにかかる時間（デッドタイム）が発生し，これにより認識しなくてはいけない音声が終わってしまうことが挙げられる（懸念点①）．

音声検知としては，音声を発している場合に音声を検知できないことにより，音声認識システムを立ち上げることができないこと（懸念点②）．また，音声以外を音声と検知してしまうことにより，無駄に音声認識システムを立ち上げ稼働時間が削減しきれないこと（懸念点③）が挙げられる．

今回検討を行った音声検知法において，ゼロクロス法では音声の特性を見るわけではないため上の懸念点②，③について考慮する必要があると考えられる．また LPC 分析では音声の特性を見るための処理に時間がかかることから，上の懸念点①について考慮する必要があると考えられる．これらの懸念点についてそれぞれシミュレーションを行った．

5．シミュレーション方法および結果

（1）ゼロクロス法（雑音下における音声検知）

音声検知を行う場合において課題となるものの 1 つとして雑音下での検知率が挙げられる．これについて，ゼロクロス法では信号が 0 を通過した回数で周波数を算出するため大きく影響を受けることが懸念される．これは第 4 章における懸念点②に当たる部分である．そのため，ここではまずゼロクロス法での音声検知において，雑音の乗った音声の *SNR* を変化させた場合，どこまで音声を検知できるかについてシミュレーションを行った．

音源には音声資源コンソーシアムのコーパスリストより，9-a.　雑音重畳日本語連続数字 音声認識評価環境および 9-b.　雑音下日本語連続数字 音声区間検出評価環境のサンプルを使用した．これらは音声に特定の雑音を *SNR* が 20 dB から-5 dB まで 5 dB 間隔で加えた音声である．それぞれの音声には雑音として（9-a）に地下鉄の音，（9-b）にレストランの音が加えられている．次の表 1，2 にシミュレーション結果を示す．今回使用した音源（9-a），（9-b）において，それぞれの音声の中心周波数は（9-a）が 290 Hz ，（9-b）が 700 Hz であることが分かっているため，この周波数を導出できるかシミュレーションを行った．

表 1　ゼロクロス法における雑音の影響（9-a）

SNR[dB]	20	15	10	5	0	-5
RMS[dB SPL]	66.5	66.6	66.6	67.2	66.8	67.9
ゼロクロス回数	14	14	18	14	20	56
周波数[Hz]	280	280	360	280	400	1120

表 2　ゼロクロス法における雑音の影響（9-b）

SNR[dB]	20	15	10	5	0	-5
RMS[dB SPL]	71.7	72.0	71.7	71.7	76.0	67.1
ゼロクロス回数	35	35	39	37	36	53
周波数[Hz]	700	700	780	740	720	1060

ゼロクロス法において，音源（9-a）に関しては表 1 から *SNR* が 20 dB，15 dB，5 dB の場合において音声を検出できていることが確認できる．また音源（9-b）に関しては表 2 から *SNR* が 20 dB，15 dB の場合において検出できていることが確認できる．

次に，音声区間外においてのシミュレーションを行った．これは第 4 章における懸念点③に当たる部分である．次の表 3，4 にシミュレーション結果を示す．この表 3，4 における *SNR* は雑音レベルが表 1，2 と同じ音源で信号がない場合であることを示している．

表 3　ゼロクロスシミュレーション（9-a）（雑音のみ）

SNR[dB]	20	15	10	5	0	-5
RMS[dB SPL]	42.4	44.1	53.4	55.2	61.9	67.9
ゼロクロス回数	66	58	74	67	62	90
周波数[Hz]	1320	1160	1480	1340	1240	1800

表 4　ゼロクロスシミュレーション（9-b）（雑音のみ）

SNR[dB]	20	15	10	5	0	-5
RMS[dB SPL]	49.0	48.0	54.7	67.5	69.5	63.2
ゼロクロス回数	40	40	29	44	38	39
周波数[Hz]	800	800	580	880	760	780

音源（9-a）に関しては，表 3 より *RMS* および周波数がどちらも人の音声の範囲と一致していないため誤検知をしていないことが確認できる．音源（9-b）に関しては，表 4 より *SNR* が 20 dB，15 dB，10 dB の場合において周波数は人の音声範囲と一致しているが，*RMS* が一致していないため誤検知をしていないことが確認できる．しかし，*SNR* が 5 dB 以下の場合では *RMS* も一致しているため人の音声と判断し誤検知を起こしていることが確認できる．

（2）LPC 分析（フレーム長に対する検知率の変化）

今回検討を行っているシステムでは第 4 章における懸念点①に当たるデッドタイムを短くすることが求められる．今回使用する候補に挙げている LPC 分析では信号処理が比較的多くなることからデットタイムが長くなることが懸念される．

ここでは音源に音声資源コンソーシアムのコーパスリストより，「20.　身体情報付き男・女・子どもの母音音声データベースのサンプル（男女各 4 サンプル）」を使用し，フレーム長を変化させていった場合における LPC 分析の音声を検知率についてシミュレーションを行い，デットタイムの短縮について検討を行った．フレーム長は音声の認識や検知を行う場合，25 ms を基準として考えられる場合が多いため，基準とされる 25 ms から 5 ms まで 5 ms 間隔フレーム長を短くしシミュレーションを行った．次の表 5 から表 9 に各母音でのシミュレーション結果を，表 10 に表 5 から表 9 で行ったシミュレーションの総合結果を示す．

表 5　フレーム長 vs. 検知率（母音「あ」）

フレーム長	男性の検知率	女性の検知率
25ms	100%	100%
20ms	100%	100%
15ms	100%	100%
10ms	100%	75%
5ms	100%	75%

表 6　フレーム長 vs. 検知率（母音「い」）

フレーム長	男性の検知率	女性の検知率
25ms	100%	100%
20ms	100%	100%
15ms	100%	100%
10ms	100%	75%
5ms	75%	75%

表 7　フレーム長 vs. 検知率（母音「う」）

フレーム長	男性の検知率	女性の検知率
25ms	100%	100%
20ms	100%	100%
15ms	100%	75%
10ms	100%	25%
5ms	75%	25%

表 8　フレーム長 vs. 検知率（母音「え」）

フレーム長	男性の検知率	女性の検知率
25ms	100%	100%
20ms	100%	100%
15ms	100%	100%
10ms	100%	100%
5ms	100%	75%

表 9　フレーム長 vs. 検知率（母音「お」）

フレーム長	男性の検知率	女性の検知率
25ms	100%	100%
20ms	100%	100%
15ms	100%	100%
10ms	100%	75%
5ms	75%	75%

表 10　フレーム長 vs. 検知率（合計）

フレーム長	男性の検知率	女性の検知率	トータルの検知率
25ms	100%	100%	100%
20ms	100%	100%	100%
15ms	100%	95%	98%
10ms	100%	70%	85%
5ms	85%	65%	75%

以上の結果より音声を検知できなくなる最初のフレーム長は男性が 5 ms で，女性が 15 ms であり，平均的にも男性に比べに女性の声の方がフレーム長を短くするにつれて音声を認識できなくなっていることが確認できる．また各母音の検知率を合計した表 10 から今回のシミュレーションにおいてフレーム長は 15 ms で 90 % 以上，10 ms でも 80 % 以上と高い検知率を出せていることが確認できる．

（3）ハードウェア（規模）の比較

ここでは各分析法の検出率のシミュレーションおよび，これを含めたハードウェア（規模）の比較を行った．（1）および（2）では，各検出法の基本状態の特性についてシミュレーションを行った．これに対して，ここでは実環境に近いシミュレーションを行う．検出率の導出には複数の音源から合計 20 箇所の音声区域について，フレーム長 25 ms でシミュレーションを行った．音源には，

音声資源コンソーシアムのコーパスリストより, 7. 重点領域研究「音声対話」対話音声コーパスのサンプルを使用した．次の表 11 に今回行ったシミュレーション結果を踏まえ，各分析法のハードウェア（規模）の比較表を示す.

表 11　各分析法のハードウェア（規模）の比較

	ゼロクロス法	LPC分析
計算処理	ゼロクロスカウント（カウンタ）	相互相関関数
		行列式計算
	RMS	開根
	絶対値	絶対値
	dBSPL	偏角
パワー	低	中
回路規模	小	中
検出率	75%	90%
演算量	0.85 MFLOPS	4.26 MFLOPS

表 11 において，FLOPS とは FLoating point number Operations Per Second の略称であり、1 秒間に浮動小数点演算が何回できるかの指標値ひいては性能値の事である。まず検出率はゼロクロス法 75 % ， LPC 分析 90 % というシミュレーション結果が得られた．この検出率は処理に必要なパワーや回路規模とトレードオフの関係である．LPC 分析では検出率が良い代わりに，行列式や開根などの処理が必要となるためパワーや回路規模がゼロクロスより大きくなってしまう．またゼロクロス法では検出率が LPC 分析と比べ悪くなる代わりに，カウンタなど軽い処理がありパワーや回路規模を抑えられる.

６．まとめ

我々は音声認識を用いた端末の低電力化に向けてマルチステップ型の音声検知システムを検討している．今回はその中でも，音声検出部に使用する音声検出法についてシミュレーションを行った.

ゼロクロス法では，第 4 節の懸念点②，③についてのシミュレーションを行った. 結果として, 音声の検知に関して *SNR* が 15 dB までは検知できることが確認できた. また，音声以外に関しては *SNR* が 5 dB 以下では音声と誤検知してしまったが，*RMS* と組み合わることで誤検知を抑制できることが確認できた.

LPC 分析では，第 4 節の懸念点①についてのシミュレーションを行った．結果としてフレーム長を基準となっている 25 ms の半分ほどまで短くした場合でも検出率が 80 % を超えることが確認できた．このシミュレーションにおいて男女で検出率が異なった理由に関しては，男性と女性の音声の周波数の違いが原因の一つとして考えられる．これにより女性の方が第 1，第 2 フォルマントが 1 つのピークにまとめられてしまったことが原因の一つとして考えられる.

我々は今回の結果から，音声検知部に使用する音声検知法としては LPC 分析の方が適していると考える．消費電力の面を見た際，処理としてはゼロクロス法の方が LPC 分析より低い．しかし，音声認識技術を用いた端末に本システムを適応することを考えた場合， LPC 分析では表 11 から 90 % 検出が可能であり，安定して適用先のシステムを正常に動作させることができると言える．これに対して，ゼロクロスでは音声認識が必要な時に 4 回に 1 回立ち上げることができないことになる．これでは適用先のシステムが本来の役割を果たせないため問題となる．また，検討中のシステムでは，位置検出技術により音声検知が立ち上がることが少なくできる．そのため，LPC 分析を用いた場合でも低消費電力で動作させることができる．以上のことから LPC 分析の方が適していると考えた．LPC 分析を用いる場合において，前述したように処理時間を考慮する必要がある．これについて，表 10 からフレーム長を 15 ms にすることで，検知精度を保ちつつ処理時間を減らせると考えられる. しかし, これでも処理が間に合わない場合も考えられる. その場合, 音声検知への入力信号を保持しておき，音声認識部を立ち上げる際に受け渡すなどの処理が必要であると考えられる.

今後の課題は現在進めている FPGA を用いたシステムの実装を行い，懸念点として挙げたものに関して比較検証を行い評価することである.

謝辞 : 本研究を進めるにあたり，多大なるご協力，ご指導を頂いた法政大学理工学部安田彰教授に多大なる感謝を申し上げます. また，同研究室の皆様にも多くご助力，助言を頂き感謝しております.

参考文献

1)荒木雅弘:フリーソフトで作る音声認識システム第 2 版, 森北出版，2017

2)相川清明, 大渕康成 : 音声音響インターフェース実践, コロナ社，2017

3) Alexandru Caruntu， Alina Nica， Gavril Toderean, Emanuel Puşchiță，and Ovidiu Buza：An Improved Method for Automatic Classification of Speech , IEEE International Conference on Automation , Quality and Testing，Robotics， Vol1， 2006

4) 荒木雅弘 : イラストで学ぶ音声認識，講談社，2015