第113回CIS研究所パートナー会議事録（一般様用）

開催日:　2021年10月1日（金）
場　所:　CIS会議室　13時～15時　＋　ZOOM 利用
特別講師:　宮永　喜一 先生
公演題:　自律型音声認識ロボット
　　ー　人工知能の活用事例　ー

ZOOM会議風景

宮永 喜一　教授•工学博士　ご経歴

昭和54年3月24日　北海道大学工学部電子工学科卒業
平成　9年4月1日　北海道大学大学院工学研究科　教授
平成28年1月1日　シドニー工科大学（オーストラリア）客員教授
平成31年4月1日　北海道大学大学院情報科学研究院　教授
平成31年4月1日　公立千歳科学技術大学　客員教授
令和2年4月1日　北海道大学名誉教授
令和2年4月1日　公立千歳科学技術大学　教授（理事•副学長•研究科長）
令和3年4月1日　公立千歳科学技術大学　理事長•学長

専門分野
　情報科学、情報通信ネットワーク、マルチメディア情報処理

本日の会議場　CIS研究所会議室
　　CIS研究所　宮永先生、寺川様、山本洋一
　　ZOOM参加　生駒様、西村様、竹内様、久米様

1) 人工知能　　ディープニューラルネットワークである。
人工知能の始まりは、裏付けの取れる論文が残っている1956年に数学者の集まりから形式論理「人間の頭の考え方を数式に載せたいということで、形式論理とか認識・認知心理学門を造り始められたものと言われている。

もっとも盛んだった時期は、1984年ごろ、人工知能第2回目のブームと言われていた時であった。　エキスパートシステムやロボット工学、自然言語処理など盛んであった。

政府（通産省）の方針により、第5世代コンピューターを創ろうと世界に発信、世界は驚いた。　これが人工知能であった。このころは、日本が世界に発信する時代であった。

最近は、2000年代に入ってニューラルネットワーク中心の人工知能　AI　になってきた。
現在、殆どの研究者の中では人口知能（AI）はディープニューラルネットワークになってきた。　しかし、エキスパートシステムやロボットも生き残って研究されている。

2) ニューラルネットワーク
ディープニューラルネットワークは何か？
もともとは人間のというか、生物の神経回路をモデル化したものである。
神経回路網の一つ一つはニューロンと呼ばれる。
　ニューロンの動きは、入力が有るかないか（電流が有るかないか）　0 or 1 に対して、重みを付けて出力する。
　この複数の出力をすべて足し合わせて
　　閾値を超えると　1 を出力する→微弱な電流が出る。
　　閾値を超えなければ　0を出力する→　電流の出力は無い。

　　これが、ニューロンの数理モデルである。

　このニューロンを幾つか層に並べて全結線する、これがニューラルネットワークで1984年代に提案されている。
このモデルは、手書き文字認識には利用されたが他の応用では成果が出ず4約40年前にブームが去ってしまった。
特に音声認識には全く役に立たなかった。

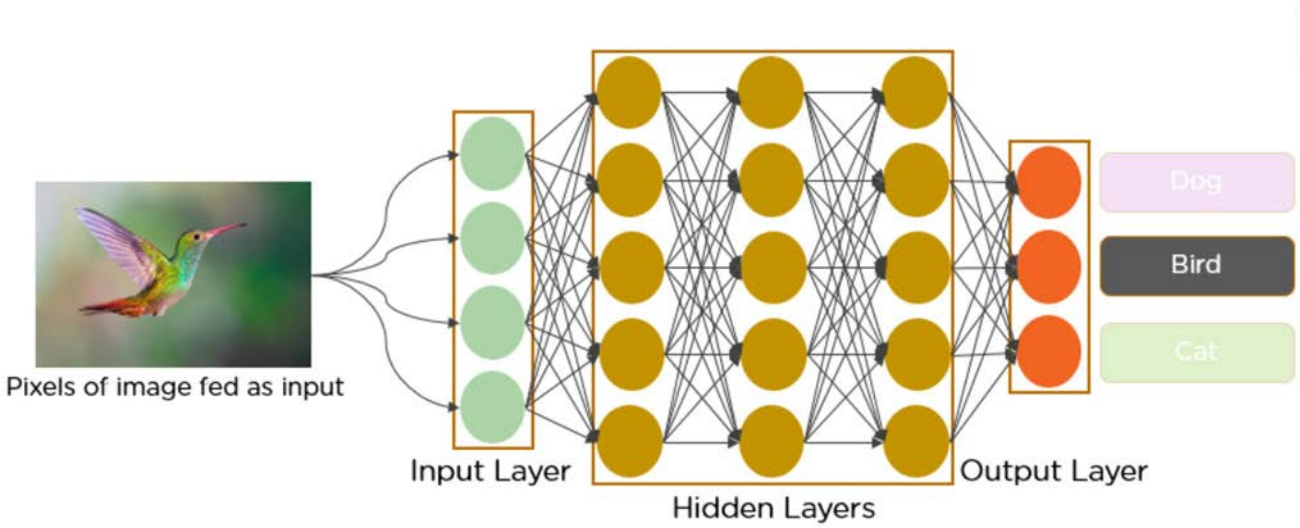
最近では、大元の画像も音声もディープ・ニューラル・ネットワーク（DNN）で直接特徴認識をする方式に代わった。これが実現できるようになったのは、40年ぐらい前のニューラルネットワーク計算能力より数千万倍の速さでの計算能力が簡単に手にはいるようになったことで可能になった。

3）Break！　―　例題　―

具体例

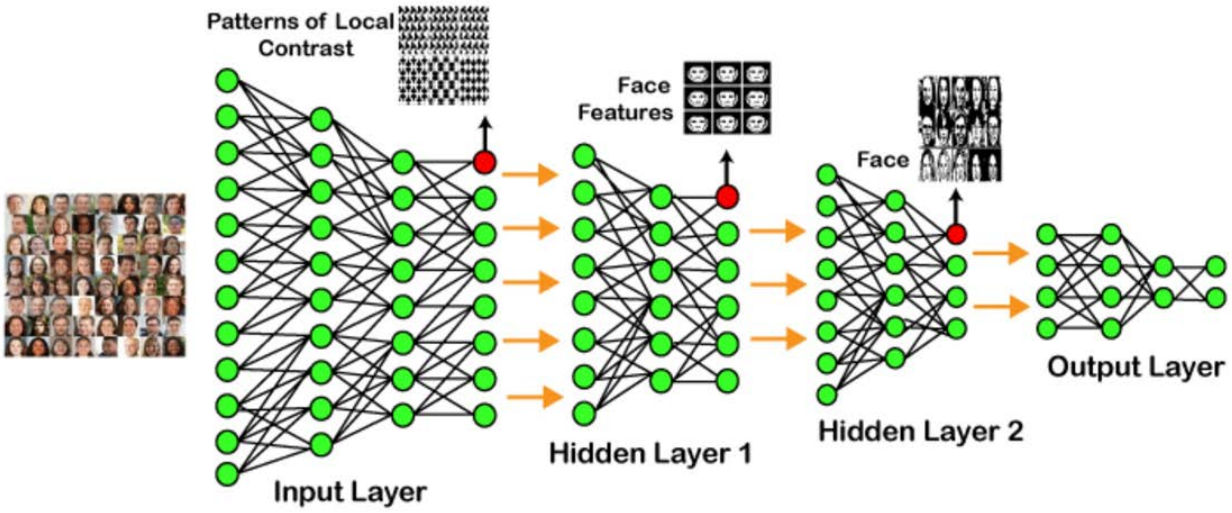
下の画像の各ピクセルデータ　例えば縦x横　1000x2000　等分して書くピクセルの*RBG* の値を一個一個入力層に与えると、そして学習済であればこの例では3層の隠れ層から3つのノードに “010”と出力されれば“鳥”であると認識されたことになる。

＊隠れ層（具体的にどういう学習したかどうかはここでは問題にしない）を通して出力層で具体的な名前（鳥）がわかる。



下図ではこれと同じように、先ず多くの画像を入れて学習させたうえで、出力は例えば名前だとすると、

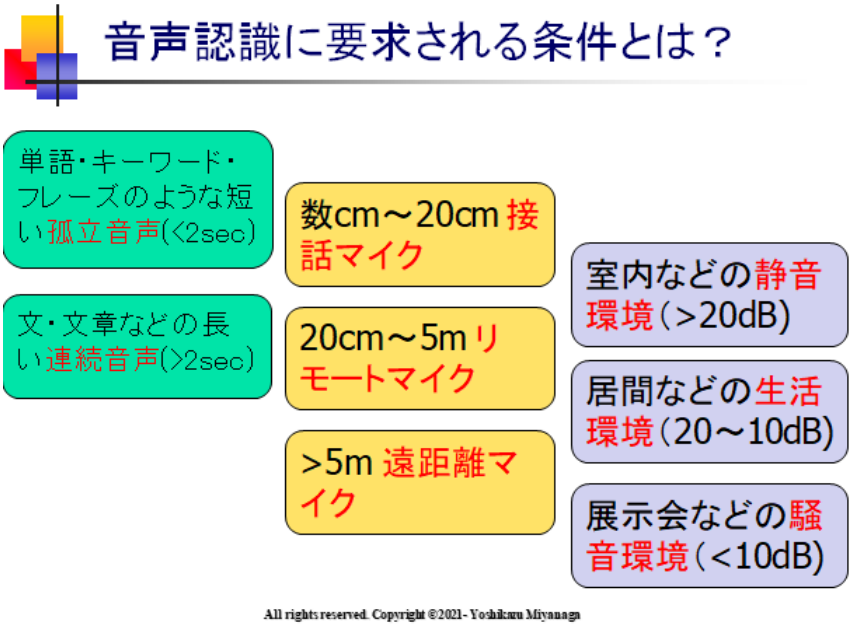
入力層では輪郭をとか、コントラストを補正しているかとかを研究している、次の隠れ層も2層がいいのか3層が良いのかなど現在は、入力層の数、下図では4層であるがこれを5層にするとか、隠れ層の数を変えるとかの最適化の研究が進められている。　これら各処理での計算能力が必要で、計算応力の進歩でDNNが進化した。



学習が終われば、ある特定の顔を入れると、その顔データーから名前が出てくる、という使われ方をしている。

DNNが良いのは、昔のニューラルネットでは顔は正面からしっかりと映っていないと認識できなかったが、今ではDNNのおかげで、横を向いていても、背景に映り込みが有っても、又顔が小さくても大きくても学習済であれば正しく認識する。　大量の学習がいるものの昔のニューラルネットワークより非常に柔軟性が高い＝汎化能力が高い。

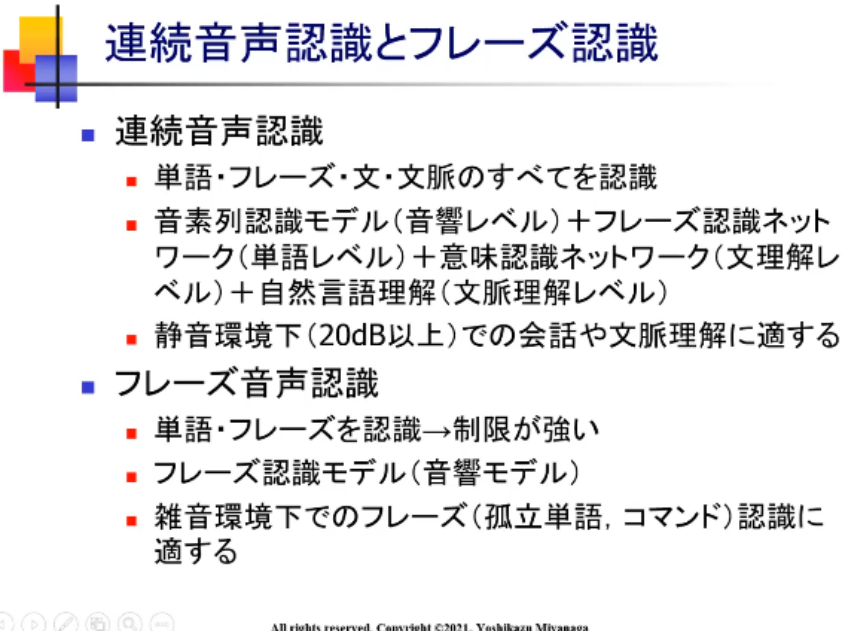
4）音声認識



音声認識は、静かな環境、うるさい環境、遠近などどのような環境でも認識する一つのアルゴリズムで全部対応できているような認識が有るが、ところが音声の”音”を考えると接話マイクの時、リモートマイク、遠距離マイクでは、音のエネルギーは距離の2乗に反比例するので急速に減衰するので信号処理は全く違ったものが必要になる。 口から数センチ離れた接話マイクはノイズの心配は全くなく雑音除去の処理は不要。 ところが、リモートマイクでは音は雑音が約0～10dBぐらいあり音声は、10dBぐらいしか取れない。 遠距離マイクでは雑音が0-10dB程度となり音声は0dB程度となってしまうのでノイズと同じレベルになってしまう。 以上の理由により、マイクの位置/場所によって処理が違ってくる。

今我々の興味はロボットにあるので、話者とは約1m程度から時としては3m程度離れていることもあり得る。この状況では、対雑音比では0～10dBである。 このような状況下、長い文章を認識するアルゴリズムだと非常に困難なので、研究室では**短い単語、キーワード、フレーズなど2秒以内の短い音声を認識するようなアルゴリズムに特化してS/N比が0～10dBでも十分認識できるようなシステムを構築**する事とした。T= 01:49:52 25:30

5）連続音声認識とフレーズ認識



長い音声は雑音環境下では認識が難しい。

基本的には3音節認識 記載例 VCV
母音(V)*子音(C)*母音(V)
母音(vowel) あ-い-う-え-お
子音(consonant)

連続音声認識は音節認識:
音節の組み合わせVCV，CVV，CV，- -を認識する（音節認識）。

長い文章の認識は、先ず音節認識をした音節から、かな漢字変換方式のような変換を行い全体の文章を認識する。

ところが、音節認識は非常に短い時間に発生された音なのでその部分に雑音が入り込むと認識がとても難しくなる。

フレーズ音声認識:
単語レベルで認識する方式。 雑音が多い場合でも、例えば、「おはよう」の最初の部分が雑音で認識できなかったとしても 「う」が認識されなかった場合でも、「は」と「よ」が認識されておれば「おはよう」と認識する。
従って雑音に強い環境＝認識が可能になる。 このアルゴリズムを採用した。

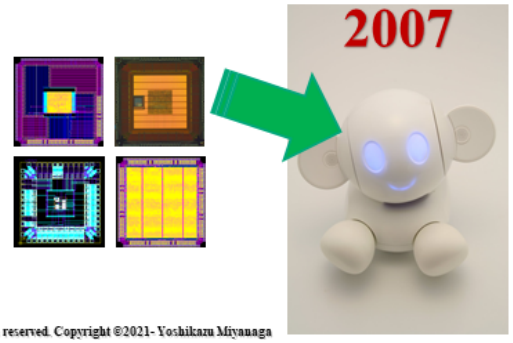
2007年大阪の会社と共同で作成:
大阪弁で会話する。
CHAPIT.mp4

登録単語は200語
　認識すると反応する。
　科学技術博物館で展示された。
　「世界で初めて人間とコミュニケーションするロボットとして展示された。

　子供がしゃべると反応しなかった。

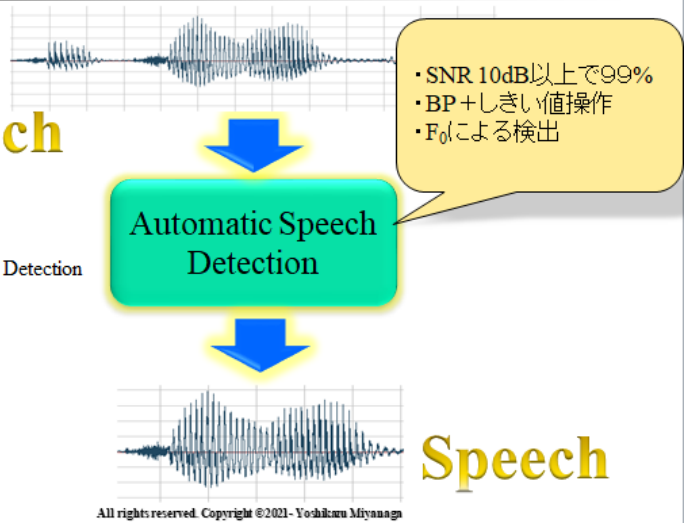
Robot Implementation

- Speech Recognition & Synthesis
- Quick Response
- Control to Consumer Electronics and Machines



マイクは常にON:
　ハローGoogle/ハローアレクサ等キーワードがなくても反応する。

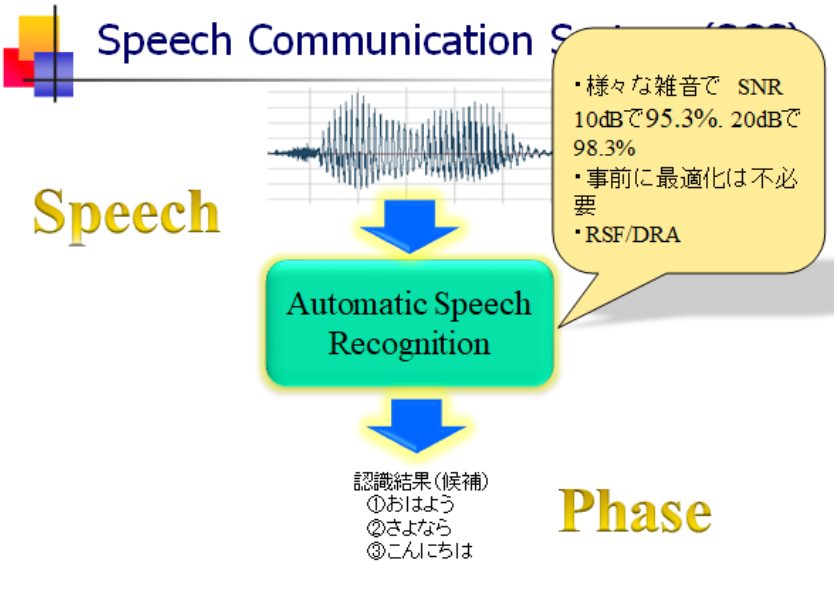
Speech Communication System (SCS)



VAD機構:
　入ってきた音が有る程度以上の　エネルギーが0.5秒程度有ると音声で有ろうと判断し、音声を抽出する。

「おはよう」
　次に、入ってきた音声は音声認識装置（Automatic Speech Recognition）で登録された単語の中から、どれが適切か調べ認識する。
第一候補の単語　「おはよう」を見つける。

Speech Communication S

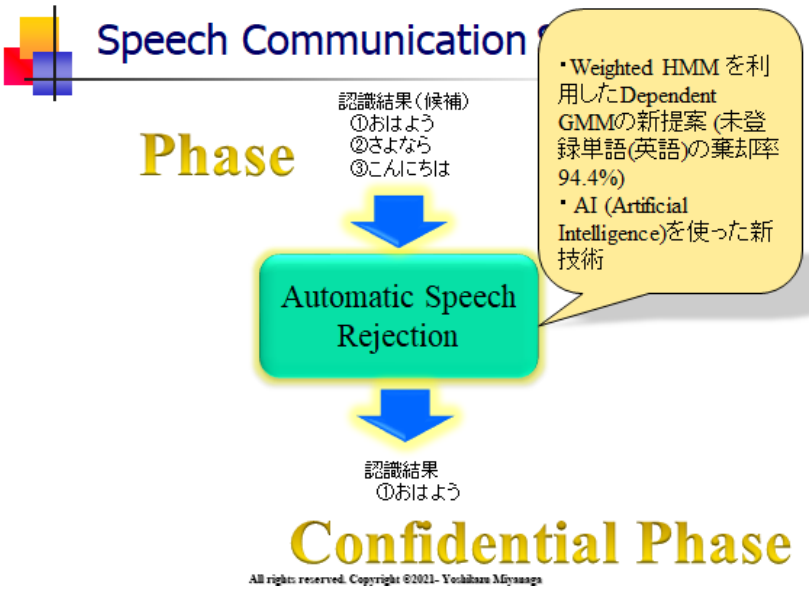


認識結果を正しいかどうか判断:

　第一候補の単語　「おはよう」が正しいかどうかチェックして正しいと思われた場合「おはよう」と判断する機構を備えた。

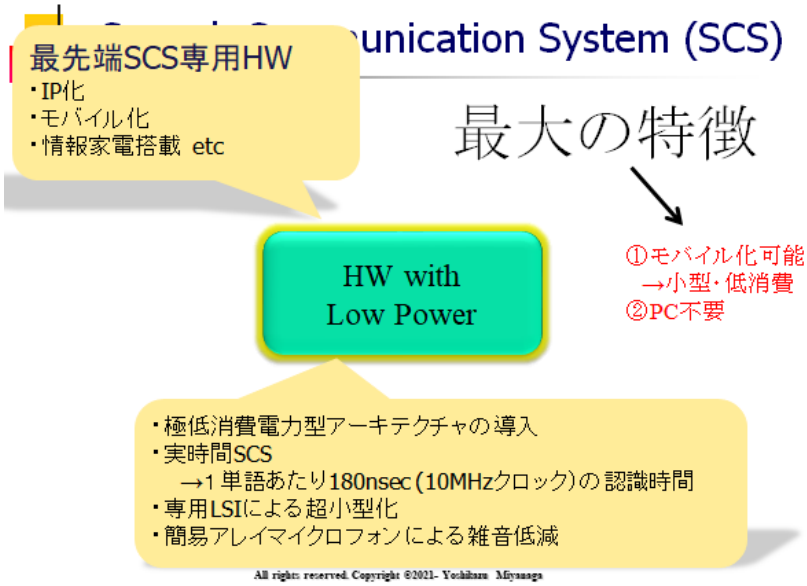
　これがない場合、ロボットは200単語しか登録されていないことから、この中に「さようなら」という単語が登録されていなかった場合、ユーザーが「さよなら」と言った場合、ロボットは理解できない。

わからないので何の反応もしない。



このロボットの最大の特徴:
一連のユニットをすべてハードウエアで作った。
Automatic Speech Detection
Automatic Speech Recognition
Automatic Speech Rejection

低消費電力
・実時間SCS
日との会話は、人間工学的に見ると
人が何か喋ったときに0.15秒以内に反応が有れば会話が成立する。　これが、0.2秒以上で反応がないと違和感を感じる。



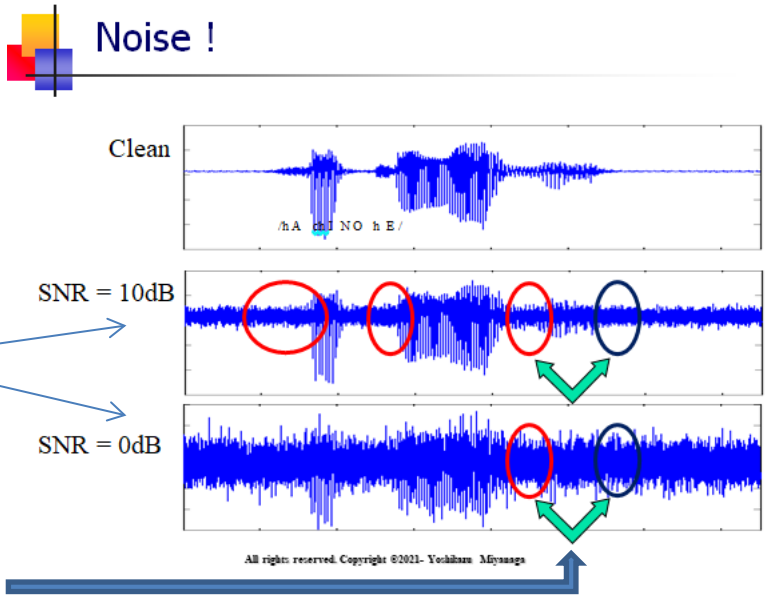
Noise！
　音声　hachinohe

　右の図で、一番上は雑音無しの（hachinohe）の波形。

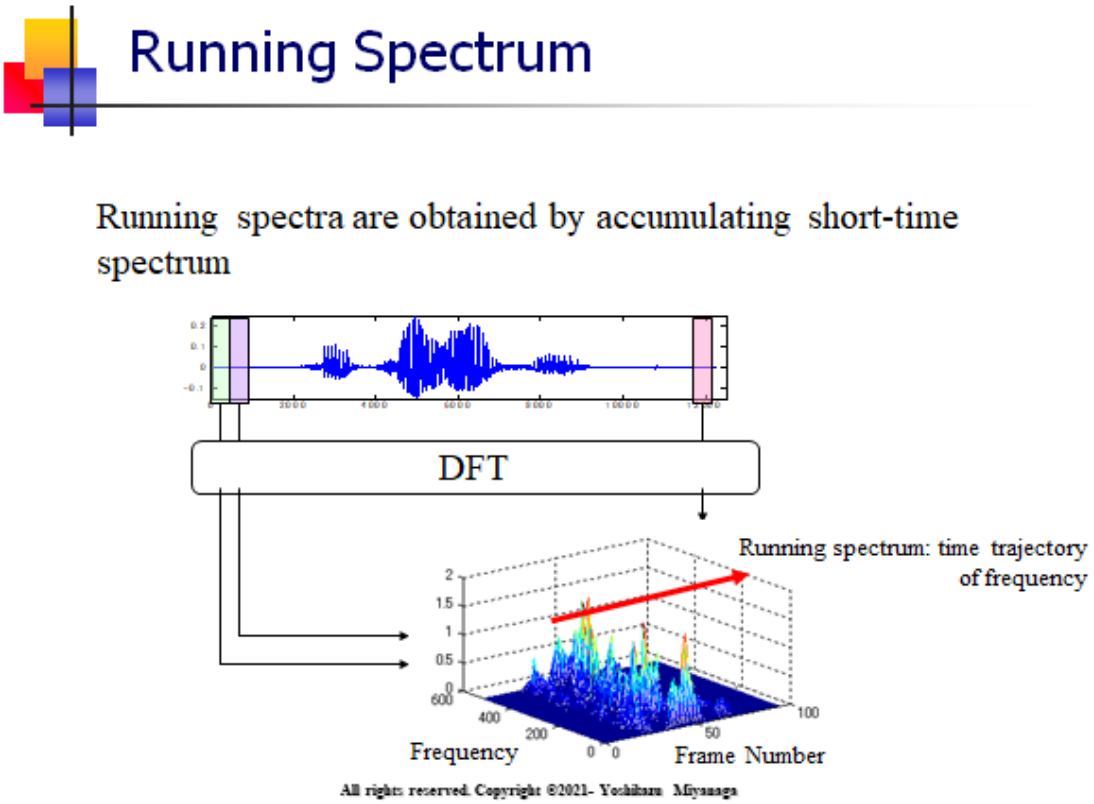
雑音比（SNR）＝10dB　（hachinohe）の波形。
後ろの赤丸は「h」　黒丸は「無音」の音

SNR = 0dB　（hachinohe）の波形。
黒丸は「無音」の音

ノイズの有る場合、赤丸「h」と音無音の区別がつかない。　この場合機械認識は不可能。
（この状態でも人は認識できるがメカニズムは分かっていない）

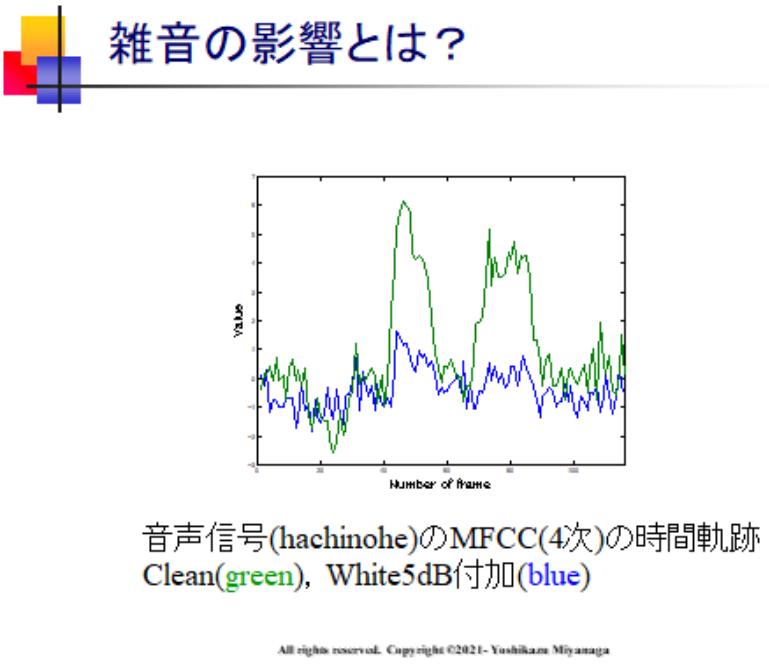


雑音除去:



上は、波形（時間軸）、10msec幅のフレームで切って、その短い時間に対してフーリエ変換する。
上の図の　周波数軸（左）ーフレーム番号（右）　に結果を示す。　フレーム番号は時間軸相当。
これでわかるように、音は時間と共に周波数特性が変化する。　　T=40:52　02:15:30

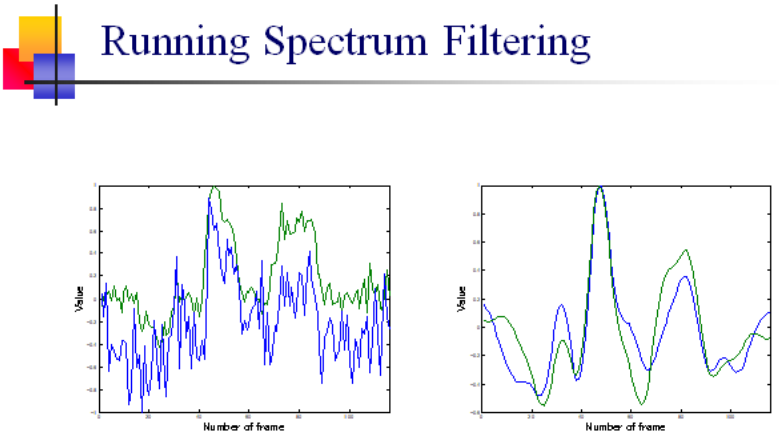
雑音の影響とは:
　雑音の影響がどれほどあるか、上の図で、周波数特性の特徴を見ると、ある固定した周波数で横軸はフレームナンバー雑音の有無で相当特徴が違う＝状況が異なっていることが分かる。



　これをどのように前処理するか（次頁に続く）

（前の図を参照しながら）例えば、エアコンや車のエンジン音などの雑音は、周波数特性は時間経過と共に一定である。　つまり、周波数特性はフレームナンバーで異なることなく一定のままである。

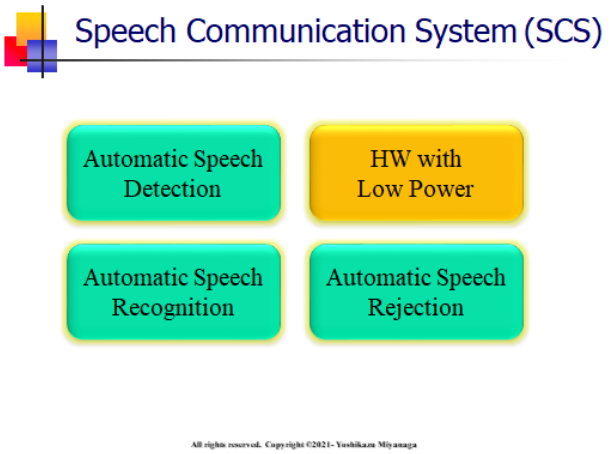
周波数を固定して、フレームナンバーに沿って（下図の赤い矢印に沿って）もう一度周波数特性をみると、定常な雑音は低くほとんど直流である。しかし、音声はころころ変わるのである程度高い周波数特性を示す。
　そこで、非常に低い周波数成分をカットするランニングスペクトル上のハイパスフィルター、正確には我々の話す声はあまり早くないことからバンドパスフィルターを適応する。



DRAのみ適用したMFCC(4次)の波形(左)と
RSF+DRAを適用したMFCC(4次)の波形(右)

上図、右はバンドパスフィルターを通した後の波形であり、雑音処理前の左側と比べて改善が認められる。
これを利用して第1世代SCSシステムを作成した。

第1世代SCSシステム:　ローパワーを目指しハードウエア化



　上図右は、ロボットに搭載したモデル（レイトロン社製）

ロボット搭載に当たっては、できる限り消費電力化が必要。

　例として、インテル社のI7のようなCPUでは、処理が速いが消費電力が非常に多きいため、バッテーリー駆動のロボットでは数分しか持たない。　低消費電力化が必要。

極低消費電力化手法:

極低消費電力化手法

- アーキテクチャ・アルゴリズムレベルからのアプローチ
  - 電源電圧V_{dd}の減少による回路遅延を補償
    - √ 近似アルゴリズムによる演算量削減
    - √ マルチレート処理
    - √ 並列処理
    - √ パイプライン処理
    - √ 処理方式の変更

→ メモリ・プロセッサ間の処理量のバランスを調整

メモリアクセスの削減

ファーストプロダクト:
シャープの掃除機に採用された。
　うらばなし
　掃除用のブラシのそばにマイクが有ったために、S/Nが悪く認識率70％で有ったが、ユーザーのアンケート調査では使いやすいと好評。　さらに、時々誤動作が有りロボットが勝手にしゃべりだすことがあった。　「今日の調子はど～」とか、突然しゃべりだす=これは完全に誤動作である。ところがユーザーからは疲れて休んでいるときに声をかけられ、とても心がやすまったという。

ココロボ と ロビ

画面、左はロビ
　ロビの誕生は、ココロボの動作をみたあるメーカーからロボットのリクエストがあり誕生した。

ROBI誕生　2013
　日本円で13万円、自分で組み立てる。
Deagostini Japan（本社イタリア）から発売。
販売実績は世界的な販路で70万台以上。

T＝48:30

　180msecで反応:700単語
　　歩く、手曲げる、踊る・・・・
　動くのに3秒かかる（メカニカルディレー）

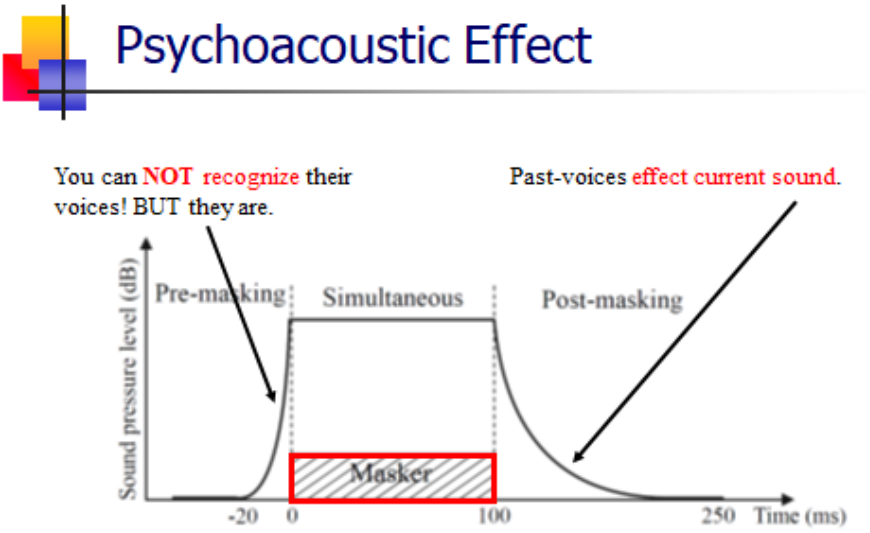
ROBI -2013-

- Producer & Sales Company
  by Deagostini Japan, and Raytron Inc, JP
- Design & Robot Controller
  by T.Takahashi, Robo-Garage Ltd
- Autonomous ASR
  by Miyanaga Lab, HU

Price = about 130K JPY

聴覚神経学:Psychoacoustic Effect　　T＝51:22

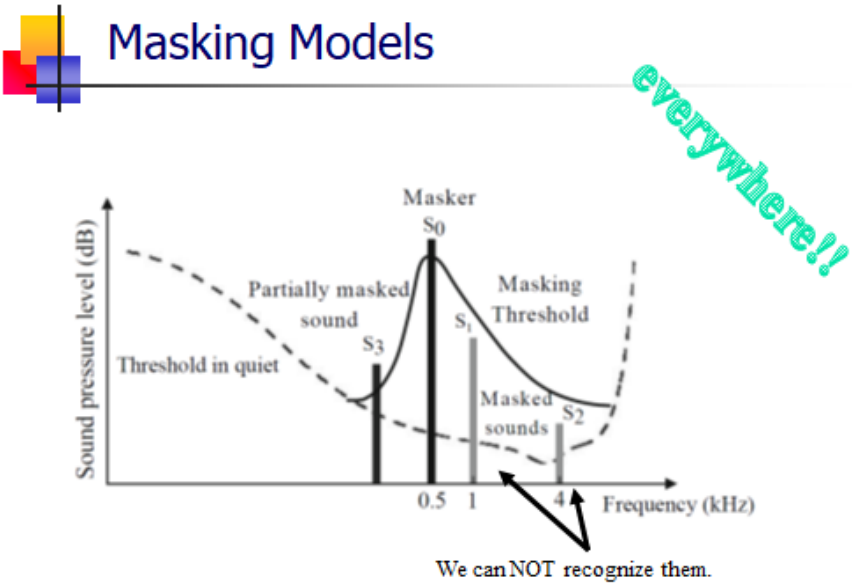
聴覚の特性がより詳細にわかりつつある。
　音を認識するときの人間のマスキング現象、静かな環境下では詳細な音が聞こえるが、うるさくなるとマスク効果が現れて、あえて聞かなくなる。　耳を保護する意味もある。



Two types of auditory masking: Simultaneous and non-Simultaneous

マスキングモデル:

　横軸は周波数
　縦軸は音のエネルギー

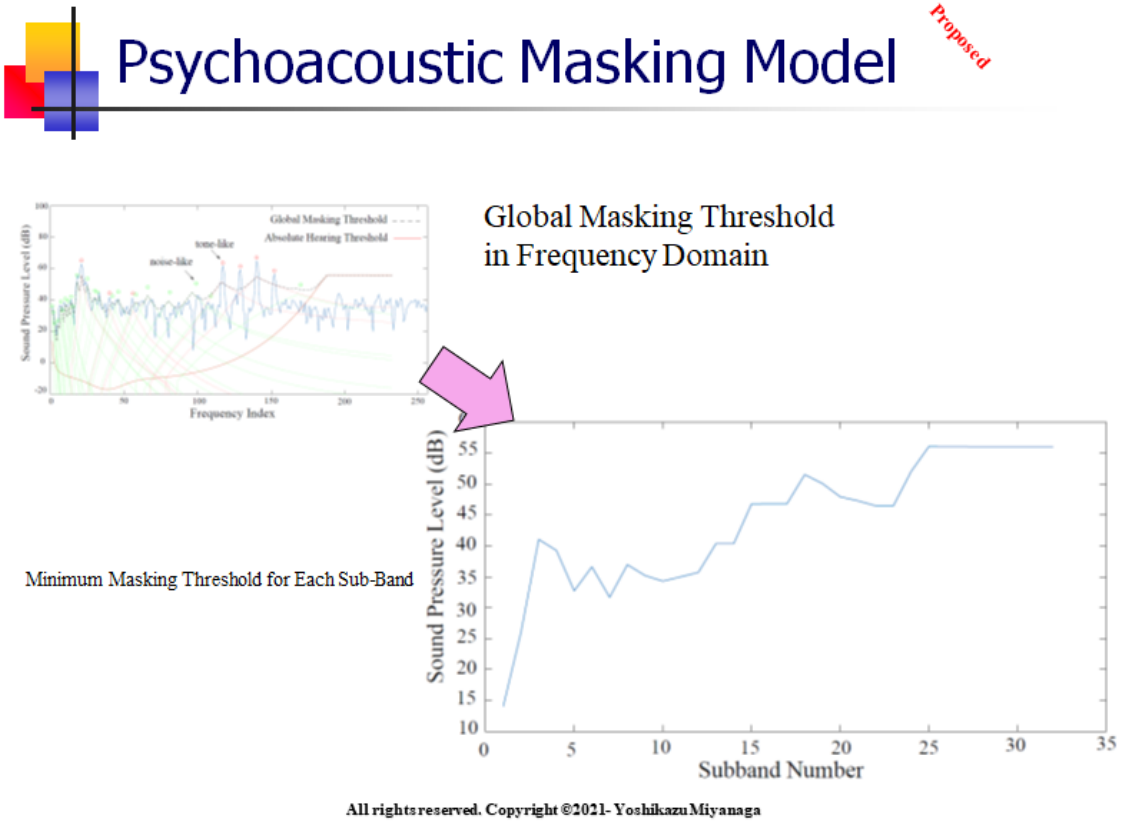


　マスクは、昔は点線がマスク現象の境界線と思われていた。

　S1 S2 S3 の単一周波数の音(ピュアトーン)を出したとすると、その音のエネルギーが点線を超えると聴こえる(超えないと聴こえない)。

　次に、二つ以上の音を混ぜて出すと、ここでは、S0 S1 S2 S3 の音を同時に出すとこの例の場合、一番音のエネルギーの大きいS0にマスク現象が実践のように引っ張られる。　その結果、実線で示された新しいマスキング域をこえるS3だけが聞こえるが、実線を超えていないS1 S2 は聞こえない。　これを、動的マスク現象(dynamic-masking phenomena)とび現在広く研究されている。

聴覚神経学のdynamic-masking phenomena　を調べた。　T=53:47

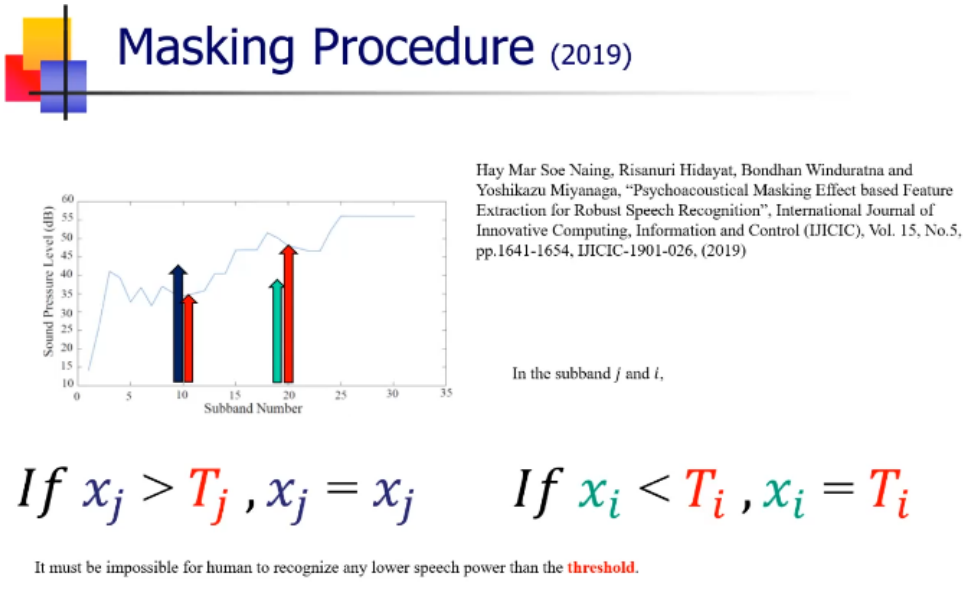


横軸のSubband Number は単一周波数でなく100Hz、200Hz毎の周波数の平均値を取ったもので周波数高い周波数はダイナミックマスクは大きく、低い周波数ではダイナミックマスクは小さい。

ということで、我々人類は沢山の音を耳で聞くときは、低い音は分解能が高くよく聞こえて、高い音はあえて聞かないようになっている、という現象が有る。

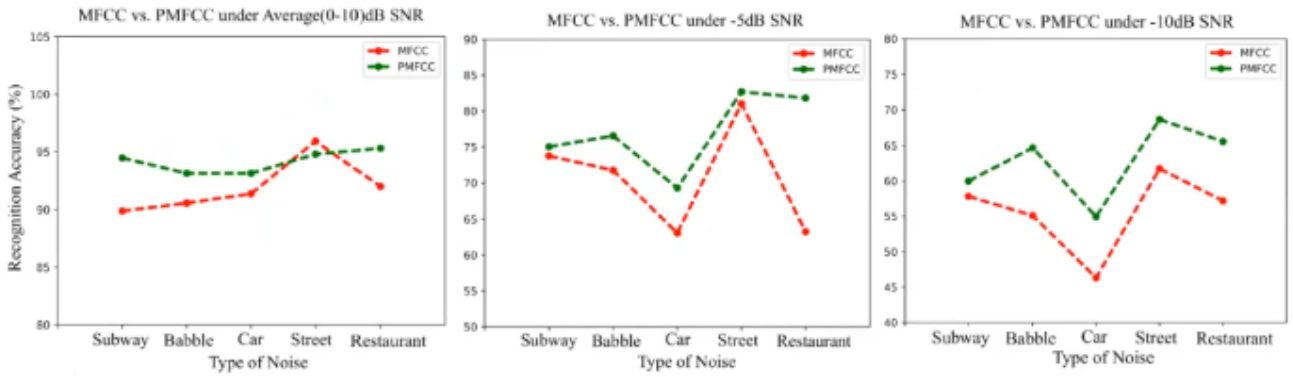
このダイナミックマスクを使って、先ずマイクに音が入ってくる、音のエネルギーを周波数解析出分析する。
計算されたエネルギーがこのダイナミックマスクより大きいエネルギーを持つ音の場合はこのまま音声認識っ装置に入れる。

ダイナミックマスクよりも小さい音＝我々には聞こえない（右図の緑）場合にはギリギリ聴こえるダイナミックマスクの壁のレベルに切り替える＝赤い信号と置き換える。
この結果、我々が聴いている状況に近い状況となる。



疑問:黒は赤に置き換えますか？
ダイナミックマスクより大きな特徴量の場合は、そのまま利用するので、置き換えはありません。そのまま利用します。

以上を実施すると、0-10dBの音、-5dB、-10dBの状況でも従来の認識装置よりは概ね性能は良い結果であった。



MFCC : conventional Mel-Frequency Cepstral Coefficients
PMFCC : Psychoacoustical masking effect-based MFCC (**Proposed**)

PMFCC can be used not only for HMM based ASR
but also CNN based ASR with D-Learning.

この実験は、データー数が非常に大きくなってDNNを使った連続音声に対する実験データーです。

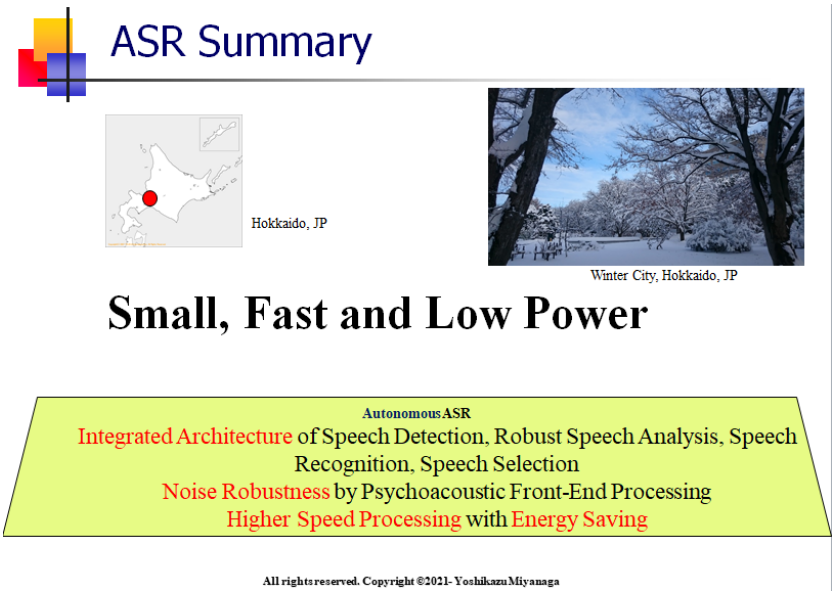
この実験での方式は、音声入力の後で、その特徴量を計算します。上記の場合は、従来のMFCCとするか、比較対象で、提案しているPMFCCとするかのどちらかです。その特徴量を、HMMモデルに入力します。ただ、この場合の認識結果は、1音節から3音節までの小さな音節単位なので、HMMから得られた認識結果をDNNに入力します。それにより、文字列認識の精度を上げています、

横軸は、いろいろな雑音減による実験で、3つの図は、SNが異なります、おおむね、提案の特徴量が良い結果を生んでいます、SNが悪いときには、その差が大きくなっています。
連続音声の音節認識で、ターゲットは100単語の認識に使ったもので、マスク現象を取り入れたほうが認識効果は高いという結果が得られた。

以上　音声認識

現在、この性能を持つ認識装置は世の中にない。
現存する音声認識は、殆どが接話マイク方式で、AIマイクもせいぜい1m程度が限界の状況である。

現在のところ、我々（北海道大学）の研究だけが雑音下でも、できるだけ人間並みに早く処理をする実績が有り研究を続けている。



完

会議風景

文責　2021-10-25　山本洋一