Speech Recognition on Mobile Devices: Distributed and Embedded Solutions

Zheng-Hua Tan[1] Miroslav Novak[2]

[1]Department of Electronic Systems
Aalborg University
zt@es.aau.dk

[2]T. J. Watson Research Center
IBM
miroslav@us.ibm.com

Interspeech 2008, Brisbane, Australia, 22-09-2008

About this tutorial

- Provide an overview of speech recognition on mobile devices
- Cover network speech recognition, distributed speech recognition and embedded speech recognition
- Presume familiarity with speech recognition fundamentals

Outline

1. Introduction
2. Network Speech Recognition
3. Distributed Speech Recognition
4. Embedded Speech Recognition
5. Applications

1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

Mobile technology

The prevalence of mobile devices: being used as digital assistants, for communication or simply for fun.

- Mobile phones: 3.5 billion by 2010
- PDAs, MP3 players, GPS devices, digital cameras

The proliferation of wireless networks: being accessible anywhere, anytime and from any devices.

- 3G, WLAN, Bluetooth, and IP networks
- Free wireless connection for the public

Mobile technology

"To the same extent that TV transformed entertainment in the 1960s and the PC transformed work during the 1980s, mobile technology is transforming the way that we will interrelate in the next decade."

- Michael Gold, SRI Consulting.

When will speech technology transform the way we interact with mobile devices, and what shall be done to make it happen?

Speech interfaces for mobile devices

Plus - opportunities:

- Advances in mobile technology: powerful embedded platforms and pervasive networking
- The course of miniaturisation
- Used while on the move
- Hands-free requirement in cars
- Navigation in complex menu structures, inevitable but beyond manageable

Minus - challenges:

- Competing with existing, well-accepted UI methods like typing on a keypad or pushing buttons
- Disturbing in public places (Remember the history of mobile phone!)
- Technical challenges

Technical challenges

Difficulty in porting state-of-the-art ASR systems onto mobile devices

- Computational constraints and power limitations
- Diverse operating systems and hardware configurations

Imperfection of networks

- Data compression
- Transmission impairments

Resources and constraints of devices

Embedded platform vs. desktop PC

	CPU	Arithmetic	RAM	Cache
HP iPAQ	624 MHz	Fixed-point	64 MB	16 KB
HP PC	3000 MHz	Floating-point	8000 MB	6000 KB

- Battery lifetime (around 3-5 h in a mobile phone when talking)
- In a consumer product, these resources are chosen according to requirements of the main functionality of the device.
  - ASR is considered but no driving forces

The targeted speech recognition application shall match the available resources, and optimization is necessary.

Resources and constraints of networks

Network availability: 'always-on' networking

- Networking facility is becoming a standard component on mobile devices
- Network service is gradually moving towards a flat-rate subscription-based business model

Network types: circuit-switched vs. packet-switched

- Circuit-switched networks
  - A dedicated circuit (or channel) btw the two parties
  - A constant delay and a constant throughput
  - Ideal for real-time communications
- Packet-switched networks
  - Routing packets through shared nodes and data links
  - Being more efficient and robust if delay is tolerable
  - To be the dominating network form (flexibility and costs)

Resources and constraints of networks

Transmission impairments:

	Landline	Wireless
Circuit-switched	reliable	bit error
Packet-switched	packet loss	packet loss

- Both bit error and packet loss tend to be burst-like, difficult to recover from

Network capacity is expanding, so are new applications. As a result, data compression is always welcomed.

- Low-bit rate compression in NSR is a source of performance degradation
- The effect of data compression on DSR is often negligible

1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

Automatic speech recognition

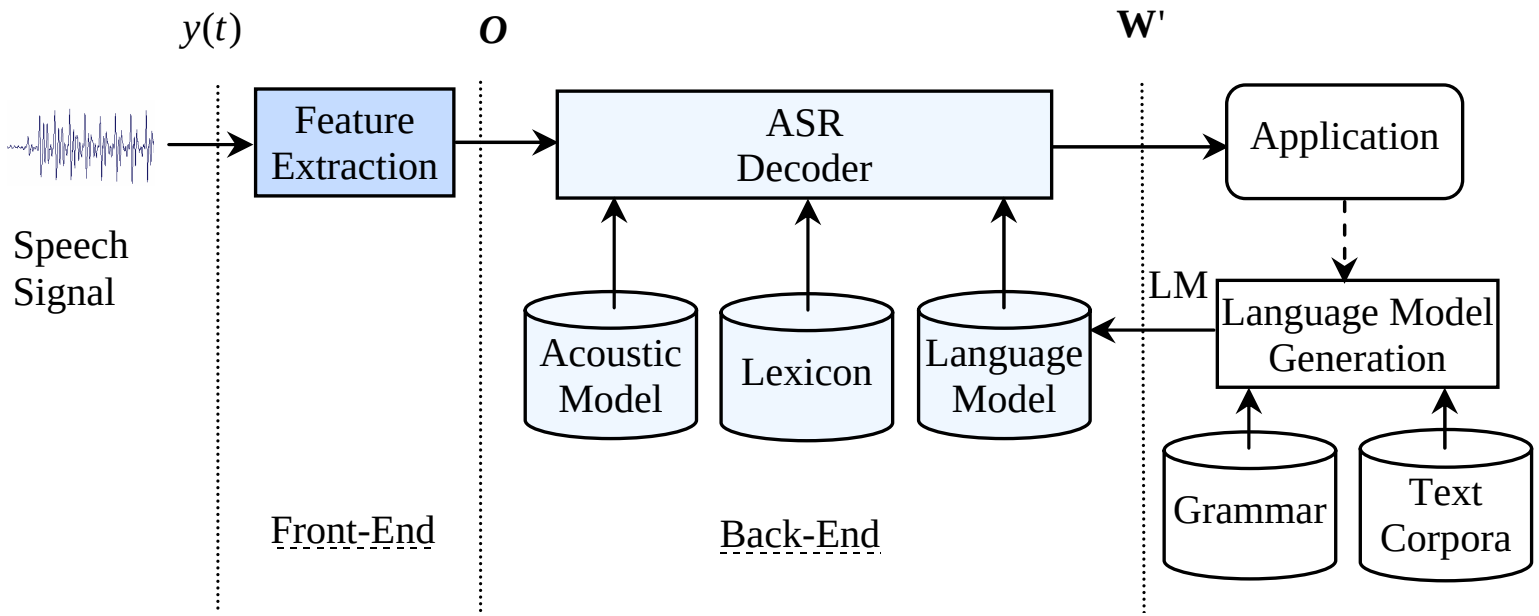
Modern ASR systems are firmly based on the principles of statistical pattern recognition, in particular the use of hidden Markov models (HMMs).

The most likely sequence of words W' is found through Bayesian decision rule:

$$\mathbf{W}' = \arg\max_{\mathbf{W}} P(\mathbf{W}|\mathbf{O}) = \arg\max_{\mathbf{W}} P(\mathbf{O}|\mathbf{W})P(\mathbf{W})$$

- $P(\mathbf{W})$ is the *a priori* probability of observing specified word sequence $\mathbf{W}$ and is given by a language model
- $P(\mathbf{O}|\mathbf{W})$ is the probability of observing speech data $\mathbf{O}$ given word sequence $\mathbf{W}$ and is determined by an acoustic model.

Architecture of an ASR System



After [Tan and Varga, 2008].

ASR components

- Feature extraction
  - Mel-frequency cepstral coefficients (MFCC)
  - Signal processing for robustness
- ASR decoding
  - Calculation of observation likelihood (based on acoustic models with millions of parameters)
  - Search (in an HMM network formed by language model, lexicon and sub-phonetic units)

Architectural solutions for ASR on devices

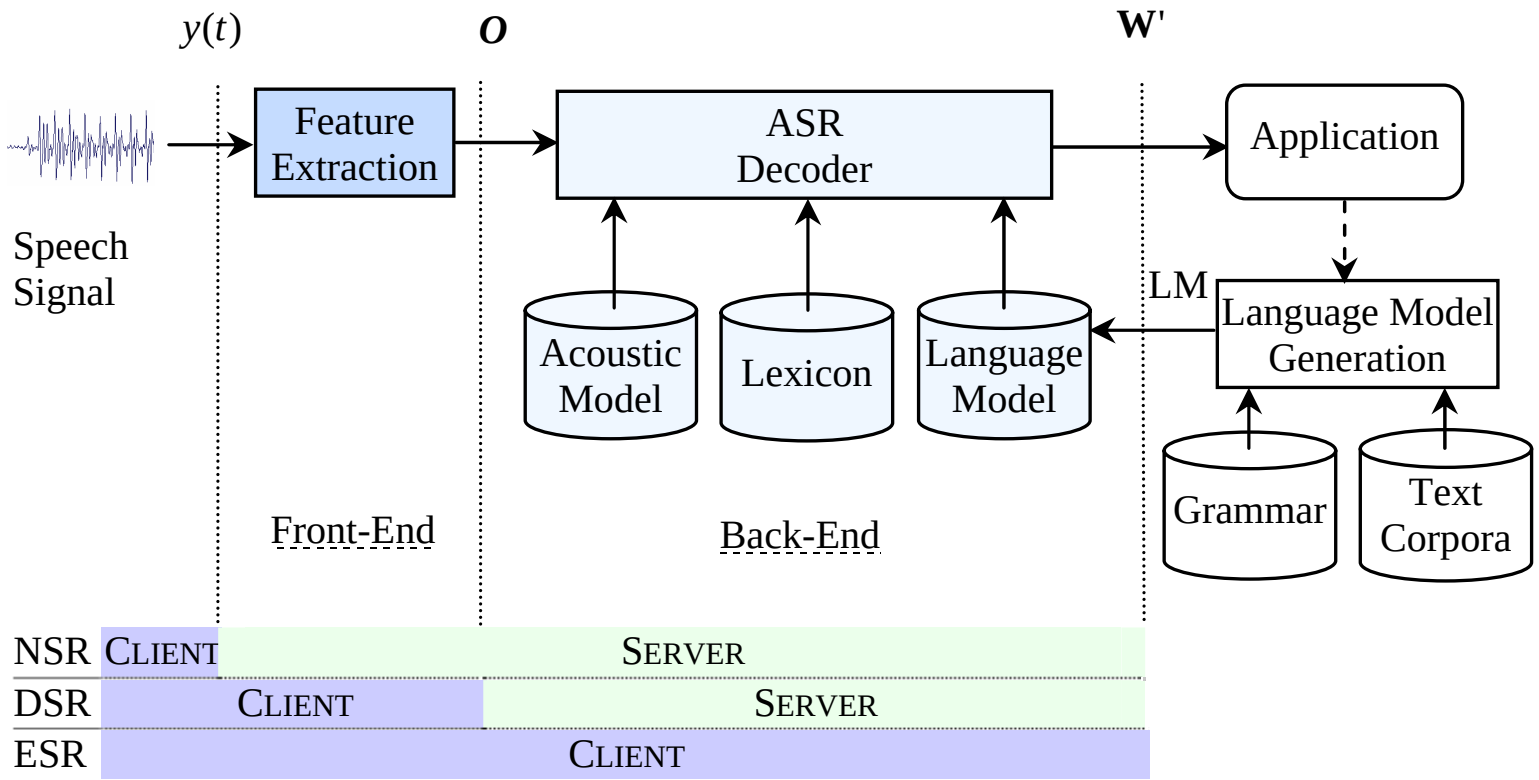
Rule of thumb for data-intensive computing is to place computation where the data is, instead of moving the data to the point of computation [Bryant, 2007].

A remote ASR may be preferable when

- The ASR requires more data from the network than from the microphone
- The ASR computation is a big burden for the device
- A quick implementation is required
- Humans assist the ASR in the background to provide semi-automatic speech transcription service

An embedded ASR may be preferable when ...

Architecture of an ASR System



The decision on where to place the ASR components distinguishes three approaches: NSR, DSR and ESR.
It is driven by factors including device and network resources, ASR components complexity and application.

1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

Network speech recognition

Network speech recognition

Remote speech recognition that uses conventional speech coders for the transmission of speech from a client device to a recognition server where feature extraction and recognition decoding take place.

Pros:

- Ubiquitous presence of codec on mobile devices
  - Plug and play, without touching the massive clients
  - The only possibility for devices like a telephone

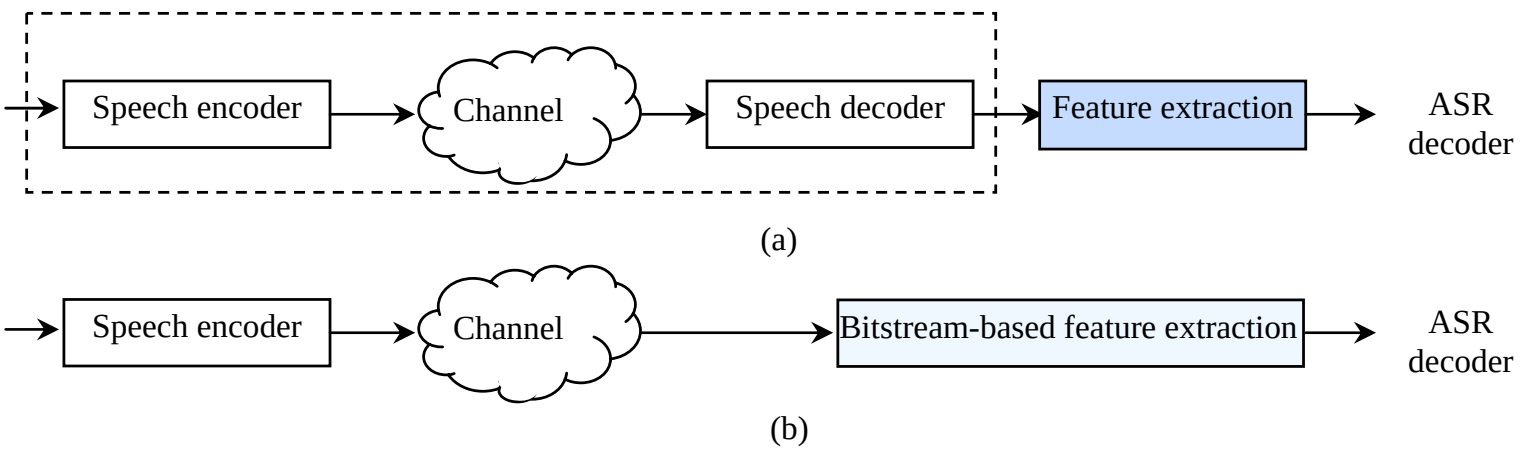
Cons:

- Network dependency and error-prone channels
  - Inter-frame dependency in coding
- Distortion introduced by low bit-rate coding
  - Linear prediction coding (LPC) vs. MFCCs

Network speech recognition

Two ways to extract ASR features from the bitstream

- (a) Reconstruction and feature extraction:
  NSR = a CODEC system + an ASR system
- (b) Feature estimation without reconstruction - bitstream-based front-end [Kim and Cox, 2001]



1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

Speech coding standards

- ITU-T:
  - G.711 PCM 64 kbps (u-law, A-law)
  - G.722.1 24 kbps, 32 kbps, 16k samples/s wideband
  - G.723.1 ACELP 5.3 kbps, 6.3 kbps (mostly in VoIP)
  - G.728 LD-CELP 16 kbps
  - G.729 CS-ACELP 8 kbps (mostly in VoIP)
- GSM:
  - GSM-FR (Full Rate) (RPE-LTP) 13 kbps
  - GSM-EFR (Enhanced Full Rate) (ACELP) 12.2 kbps
- 3GPP:
  - AMR-NB 4.75-12.2 kbps
  - AMR-WB 6.6-23.85 kbps
- IS-136 TDMA
  - IS-641 ACELP 7.4 kbps

Effect of speech coding on ASR performance

Tourist info task (5kw vocab) [Besacier et al., 2001]

	WER%
MPEG Lay2 64 kbps	7.5
None	7.7
MPEG Lay3 64 kbps	7.8
G.711 64 kbps	8.1
G.723.1 5.3 kbps	8.8
MPEG Lay1 32 kbps	27.0
MPEG Lay3 8 kbps	66.2
MPEG Lay2 8 kbps	93.8

Connected digit recognition [Kim and Cox, 2001]

	WER%
Wireline ASR	3.83
IS-641	5.25
Bitstream-based	3.76

Effect of speech coding on ASR performance

Aurora 2 database is the TI digit database artificially distorted by adding noise and using a simulated channel distortion [Hirsch and Pearce, 2000].

WER% for Aurora 2 when training and testing recognizer in the same coding mode [Hirsch, 2002]

PCM	26.77
GSM-EFR	28.56
AMR475	29.84
ALAW	29.85
AMR102	31.62
GSM-FR	31.69
GSM-HR	33.56

GSM-EFR performs the best among the codecs.

1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

Effect of transmission error on ASR performance

Aurora 2 database (clean speech only, baseline WER% being 1.77) [Kiss, 2000]

	Error-free	EP1	EP2	EP3
GSM-EFR	2.53	3.02	4.35	12.87
DSR	2.01	2.01	2.06	8.98

Network speech recognition

- Supports a wide range of devices in a plug and play fashion
- Has low requirement for the client devices
- Suffers from coding distortion, especially when it is coupled with transmission errors
- Suffers from transcoding distortion in heterogeneous networks

1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

Distributed speech recognition

Distributed speech recognition

Remote speech recognition that adopts the client-server architecture by placing the feature extraction in the client and the computation-intensive recognition decoding in the server.

Pros:

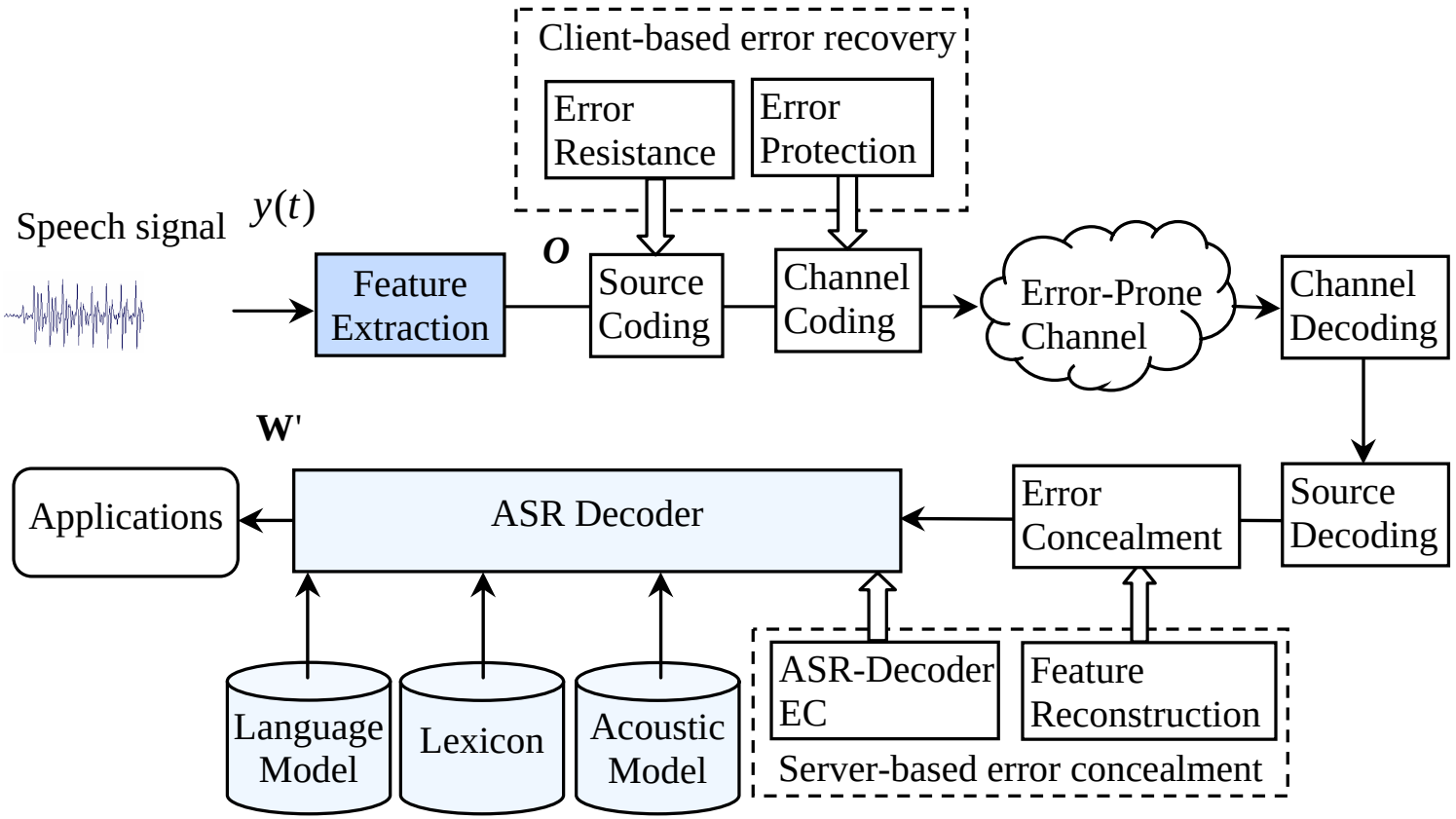
- The absence of coding and transcoding problems
- Robustness against comm. channel & acoustic noise
- Thin client, easy to update, no limits in ASR complexity
- Sever-side playback, semi-automatic transcription
- Speech data collection for AM/LM adaptation (like search engines)

Cons:

- Front-end must be implemented in the device

  (Not an issue if the application requires a client-side installation anyway.)
- Network dependency and transmission errors

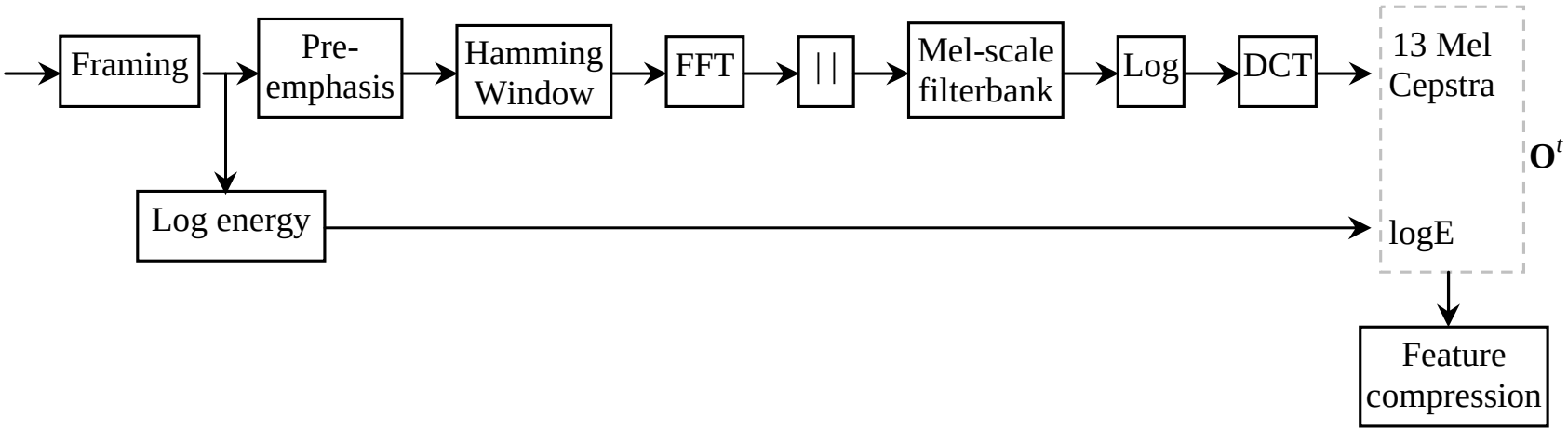
Architecture of a DSR System



From [Tan and Varga, 2008].

ETSI STQ-Aurora DSR front-end

Mel-cepstrum front-end and compression [ES 201 108]:



$$O^t = [(c_1^t, c_2^t), \ldots, (c_{11}^t, c_{12}^t), (c_0^t, \log E^t)]^T$$

Feature-pair

$$= [\,[S_0^t]^T, \quad \ldots, \quad [S_5^t]^T, \quad [S_6^t]^T\,]^T$$

Subvector1
6 bits

Subvector2
8 bits

44 bits

ETSI STQ-Aurora DSR front-end

Frame-pair architecture

Frame 1	Frame 2	CRC 1-2	...	CRC 23-24
<44 bits >	<44 bits>	<4 bits>	...	<4 bits>

< 138 octets / 1104 bits > for 12 frame-pairs

Multiframe

Sync Seq	Header	Frame packet
2 octets	4 octets	138 octets

< 144 octets / 1152 bits > for 240 ms

Bitrate

4.8 kbps with a payload of 4.4 kbps

DSR processing

The objectives of DSR processing are to achieve

- Low bandwidth requirement
- High error-robustness
- Low complexity and delay

DSR processing is all about redundancy:

- Source coding: reduce redundancy
- Channel coding: add redundancy
- Error concealment: exploit redundancy

1 Introduction
- Devices and networks
- Automatic speech recognition

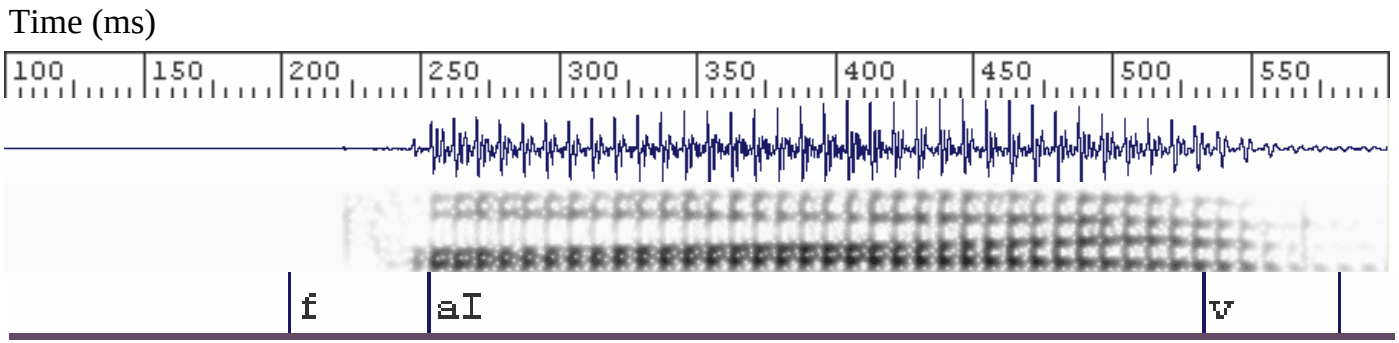
2 Network Speech Recognition
- Speech coding
- Transmission errors

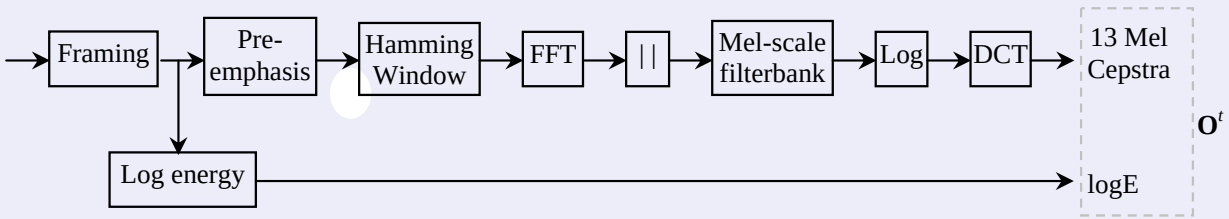
3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

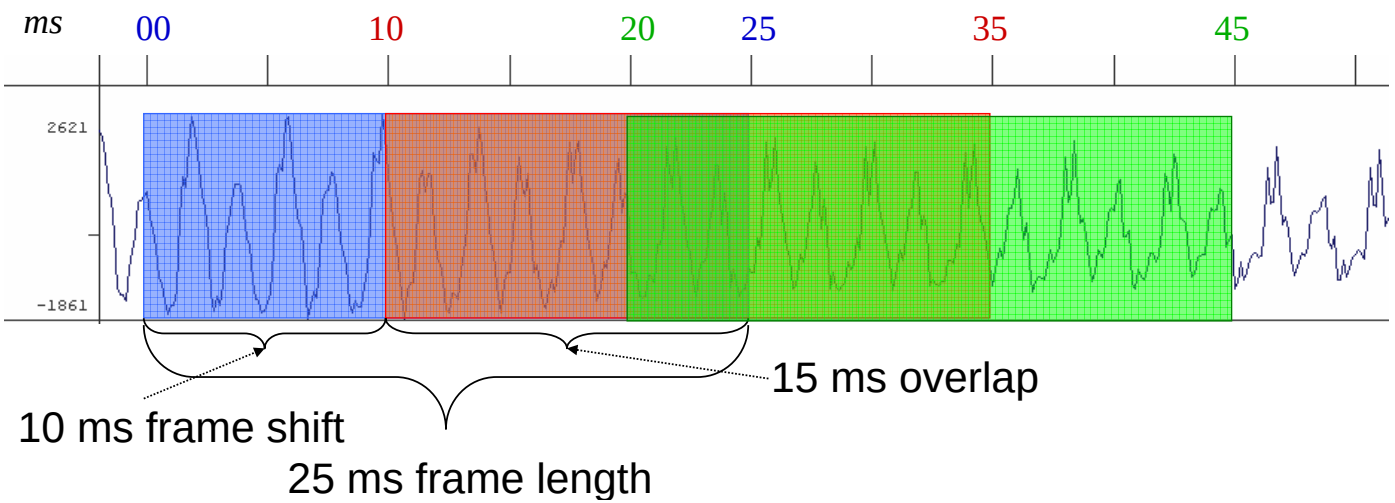
4 Embedded Speech Recognition

5 Applications

Redundancy in speech features





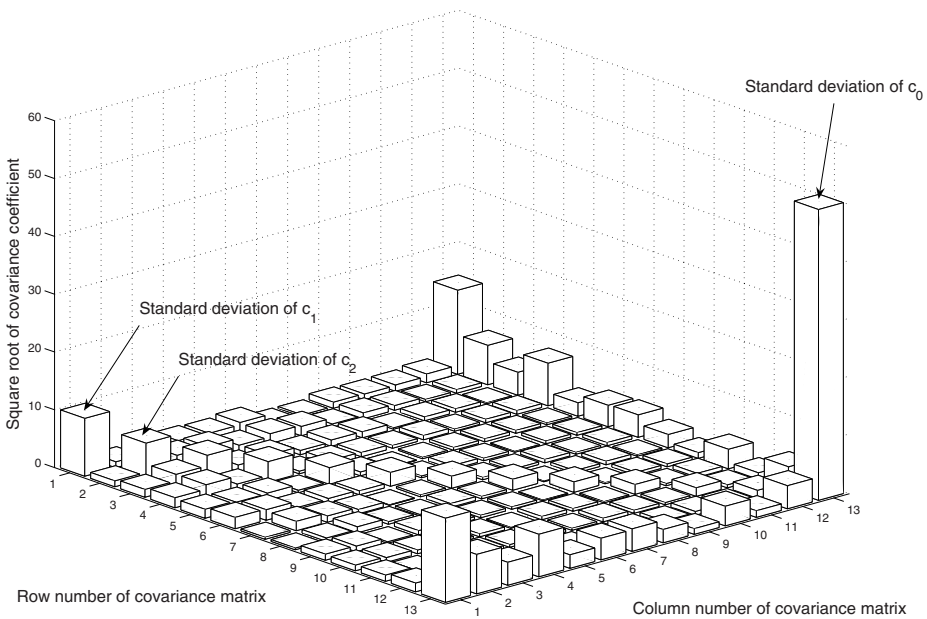


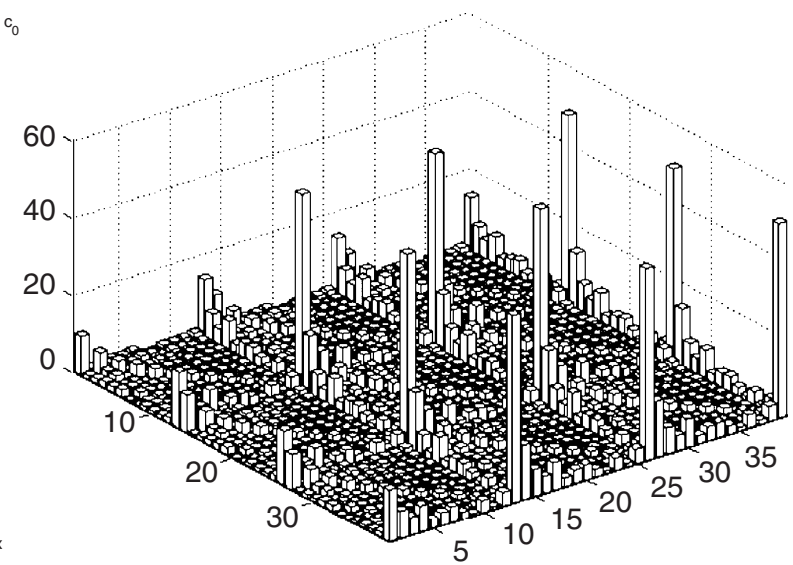
Correlation within and across MFCC Vectors

Temporal correlation (redundancy) in feature stream due to

- The overlapping in feature extraction processing
- The speech production process itself

Correlation within and across MFCC vectors (from [So and Paliwal, 2008]):





1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

Source coding

Source coding is to compress information for transmission over bandwidth-limited channels.

Transmission of uncoded feature vectors requires a bitrate of 41.6 kbps

- 13 MFCCs, 100 Hz frame rate and 32 bit floating point value

State-of-the-art DSR quantization techniques can achieve a bitrate of 300 bps [So and Paliwal, 2008].

Quantization is a process of lossy coding with the challenge being the rate-distortion trade-off.

Quantization

- Scalar quantization (SQ): input samples are quantized individually
- Vector quantization (VQ): input samples are quantized as vectors [Digalakis et al., 1999]

  Split VQ: each vector is partitioned into subvectors which are then independently quantized, as done in the DSR front-end: $\mathbf{O}^t = [[\mathbf{S}_0^t]^T, ...[\mathbf{S}_6^t]^T]^T$
  - Lower storage and computational requirement than full VQ
  - Significantly better performance than SQ at any bit-rate

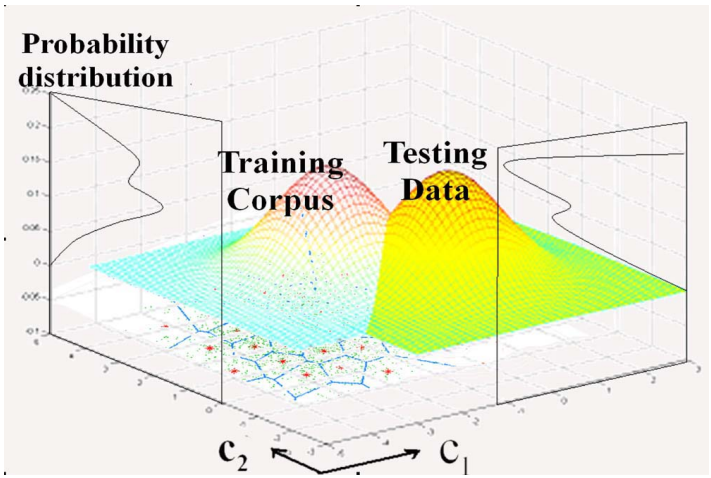
- Block quantization (transform coding)

Quantization

- Scalar quantization (SQ)
- Vector quantization (VQ)
- Block quantization (transform coding): the components of a block of samples are decorrelated by using a linear transformation (eg DCT, PCA) before SQ
  - 2D-DCT [Zhu and Alwan, 2001]
  - GMM-based block quantization [So and Paliwal, 2006]
  - Efficient but with drawbacks:
    Inter-frame coding exploits correlation across consecutive MFCC vectors, so error in one frame has considerable impact on the quality of the following frames.

Histogram-based quantization

Acoustic noise may move feature vectors to a different quantization cell in a fixed VQ codebook, introducing extra distortion!

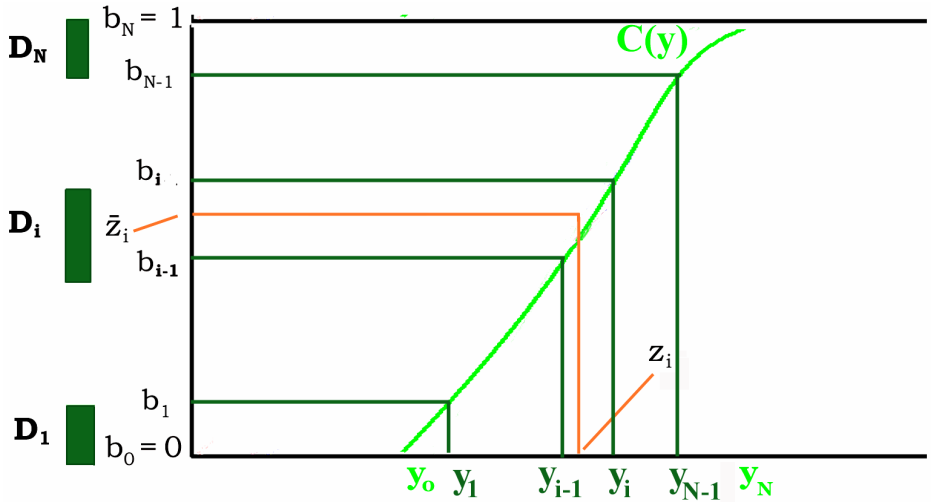


From [Wan and Lee, 2008].

HQ: The partition cells are dynamically defined by the histogram of a segment of the most recent past values of the parameter to be quantized.

Histogram-based quantization

A dynamic quantization, based on signal local statistics, not on any distance measure, nor related to any pretrained codebook.



Aurora2 (SetA,B,C) (WER%) From [Wan and Lee, 2008].

MFCC	SVQ 4.4k	2DDCT 1.45k	HVQ1.9k	HQ3.9k
38.92	43.49	40.11	22.76	18.74

HQ is also better than methods like MVA, PCA and HEQ.

Source coding & error-resistance

A low bit-rate source coding method is highly sensitive to transmission errors.

There is a trade-off between the error-resistance and the low bit-rate achieved by the removal of redundancy.

No free lunch theorem

Coding efficiency multiplied by robustness is constant.
[Ho, 1999]

So, error recovery and concealment has a role to play ...

1 Introduction
- Devices and networks
- Automatic speech recognition

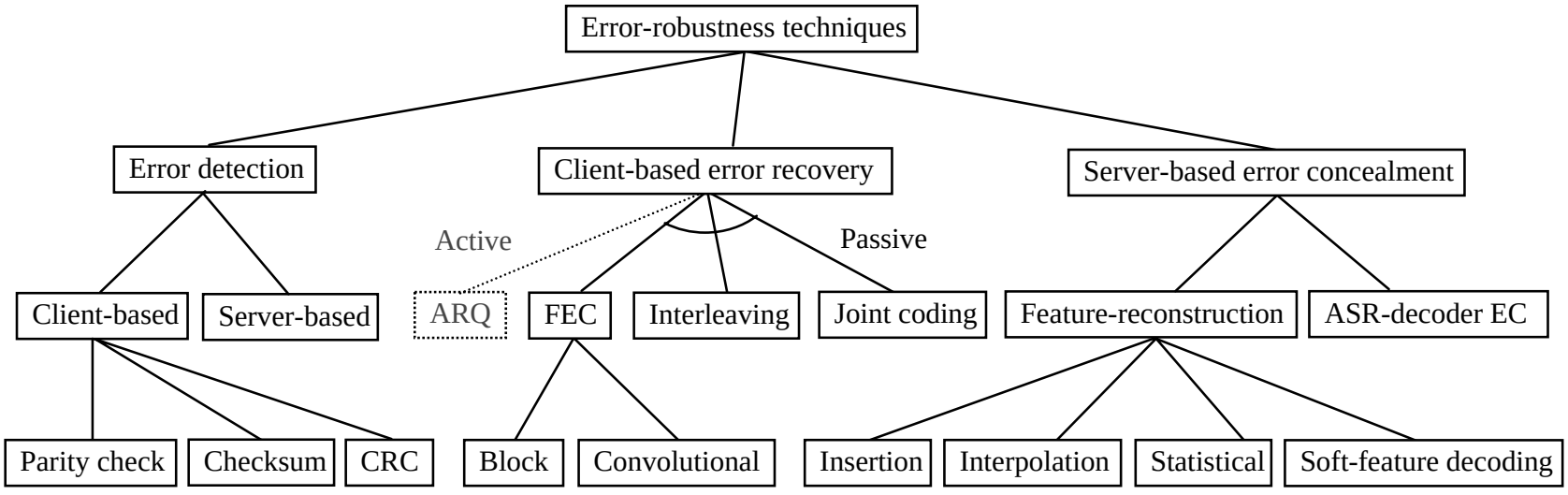
2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

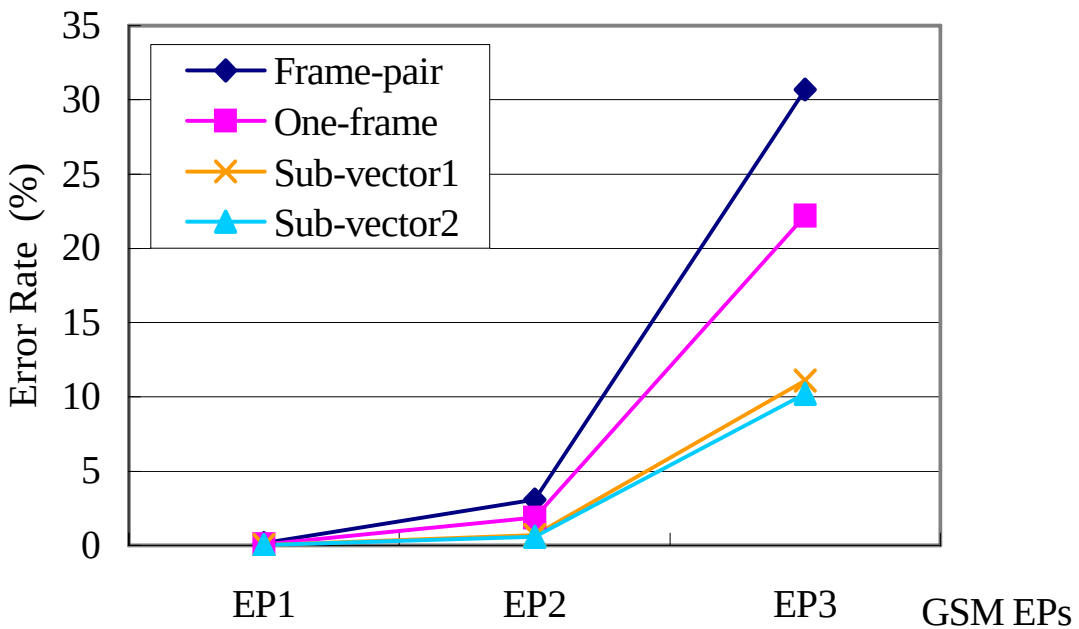
Error-robustness techniques



From [Tan et al., 2005].

Error detection

- Error detection methods
  - CRC (cyclic redundancy check), linear block codes
  - consistancy test
- Data block size



Error recovery - client based techniques

- Channel coding:
  - Forward error correction (FEC) [Borgstrom et al., 2008]
    - media-specific FEC
    - media-independent FEC: e.g. (n, k) block encoding (Reed-Solomon, BCH, Golay)
  - Multiple description coding (MDC): encoding a source into 2+ substreams to be delivered on separate channels
  - Joint source and channel coding: UEP (unequal error protection)
- Packetization
  - Interleaving: to counteract burst errors at the cost of delay [Milner and James, 2006]

Error recovery - client based techniques

A common attribute is the participation of the client aimed at exploiting the characteristics of channels and signals.

It is always a trade-off btw the achieved performance and the required resources:

- FEC trades bandwidth for redundancy
- MDC trades multiple channels for uncorrelated transmission errors among descriptions
- Interleaving trades delay for randomizing error distribution.

One disadvantage is their weak compatibility.

Error concealment - server based techniques

EC generally deploys the strong temporal correlation residing in speech features and uses the statistical info about speech.

EC techniques

- Feature-reconstruction EC: create a substitution as close to the original as possible.
- ASR-decoder EC: modify ASR decoder to handle degradations introduced by transmission errors - unique to DSR

Error concealment - server based techniques

Feature-reconstruction EC:

- Insertion-based techniques: splicing, mean value substitution, repetition
- Interpolation-based techniques: linear, cubic
- Soft-feature decoding based techniques [Peinado et al., 2003]
- Statistical-based techniques: use a priori info about speech features [Gomez et al., 2003]

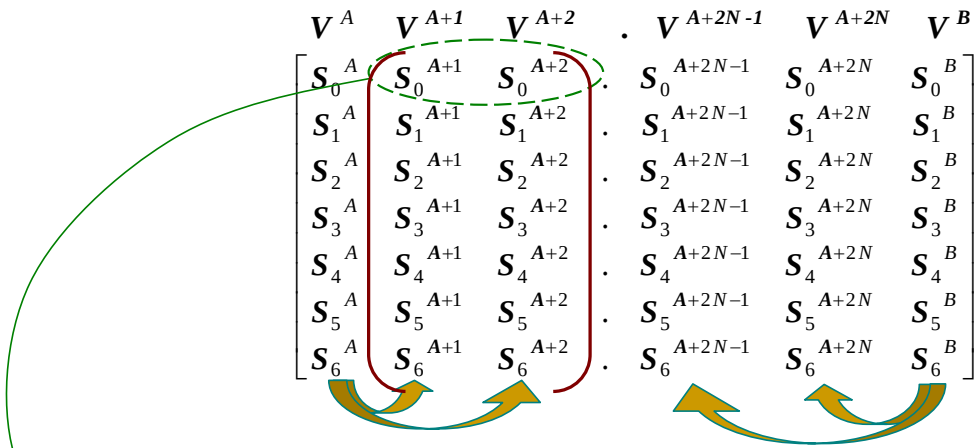
ASR-decoder EC:

- Weighted Viterbi decoding [Cardenal-Lopez et al., 2004], [Tan et al., 2007]
- Uncertainty decoding [Ion and Haeb-Umbach, 2006]

Repetition EC at subvector level

EC generally operates at vector level, yet error rates for subvector are significantly lower [Tan et al., 2007].

Buffering matrix



Consistency test

$$(d(S_j^{t+1}(0) - S_j^{t}(0)) > T_j(0)) \, OR \, (d(S_j^{t+1}(1) - S_j^{t}(1)) > T_j(1))$$

$C =$

	V^A	V^{A+1}	V^{A+2}	V^{A+3}	V^{A+4}	V^{A+5}	V^{A+6}	V^{A+7}	V^{A+8}	V^B
S_0	1	1	1	0	0	0	0	1	1	1
S_1	1	1	1	1	1	0	0	1	1	1
S_2	1	1	1	1	1	1	1	1	1	1
S_3	1	0	0	1	1	1	1	1	1	1
S_4	1	1	1	1	1	1	1	0	0	1
S_5	1	1	1	1	1	1	1	1	1	1
S_6	1	0	0	1	1	0	0	1	1	1

0 for inconsistent
1 for consistent

Weighted Viterbi decoding

Weighted Viterbi decoding

$$\delta_t(j) = \max_i [\delta_{t-1}(i) a_{ij}][b_j(\mathbf{O}^t)]^{\gamma(t)}$$

$$\gamma(t) = \begin{cases} \alpha^n, & n = 1...N/2 \\ 1 - \alpha^{N-n+1}, & n = N/2+1...N \end{cases}$$

Feature-based weighted Viterbi decoding

$$\delta_t(j) = \max_i [\delta_{t-1}(i) a_{ij}] \prod_{k=1}^{K} [b_j(o^t(k))]^{\gamma_k(t)}$$

$$\gamma_k(t) = \begin{cases} \alpha^{d(o^t(k), o^{t+1}(k))/T_k}, & S_j^t \text{ consistent} \\ \gamma_k(t+p).\beta^{|p|}, & o^t(k) \text{ substituted by } o^{t+p}(k) \end{cases}$$

Uncertainty decoding

In standard form, state emission prob. (modelled by GMM) is

$$b_j(\mathbf{O}^t) = p(\mathbf{O}^t|s_j) = \sum_{k=0}^{K-1} w_{jk} N(\mathbf{O}^t; \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk})$$

where $\mathbf{O}^t$ is the observing vector, and s_j is the state.

In uncertainty decoding, $\mathbf{O}^t$ is considered corrupted and the uncorrupted, unobservable vector $\mathbf{X}$ is a random variable with a distribution $p(\mathbf{X}|\mathbf{O}^t)$.

Integration over the feature uncertainty:

$$b_j(\mathbf{O}^t) = \int p(\mathbf{X}|\mathbf{O}^t) b_j(\mathbf{X}) d\mathbf{X} = \sum_{k=0}^{K-1} w_{jk} N(\boldsymbol{\mu}_{\mathbf{X}|\mathbf{O}^t}; \boldsymbol{\mu}_{jk}, \boldsymbol{\Sigma}_{jk} + \boldsymbol{\Sigma}_{\mathbf{X}|\mathbf{O}^t})$$

The standard HMM decoding remains, but the variance of each Gaussian is increased.

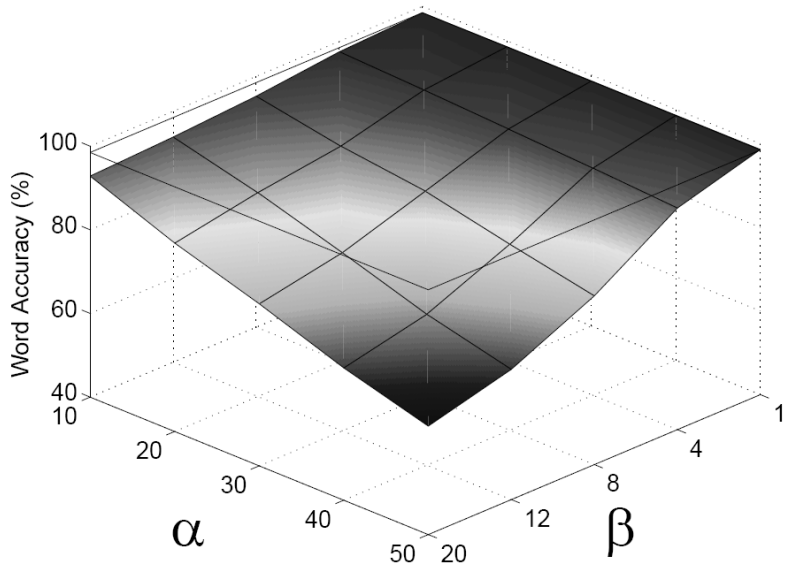
Error concealment - server based techniques

Remarks:

- No requirement for modifying the client-side of DSR, compatible with the ETSI-DSR standards
- Repetition EC works pretty well with short burst length
- Statistical based techniques benefit from *a priori* knowledge of speech and is useful in particular when burst length is long
- ASR-decoder based techniques are unique for DSR and can be applied in combination with other EC
- Computational cost is of concern

A frame-rate perspective

- Strong temporal correlation in speech features
- ASR performance is intact with a frame loss rate (short burst-length) of 50% (From [James and Milner, 2004])



So why not deliberately drop some speech frames (e.g. applying HFR, VFR), and then conducting repetition based "error concealment"?

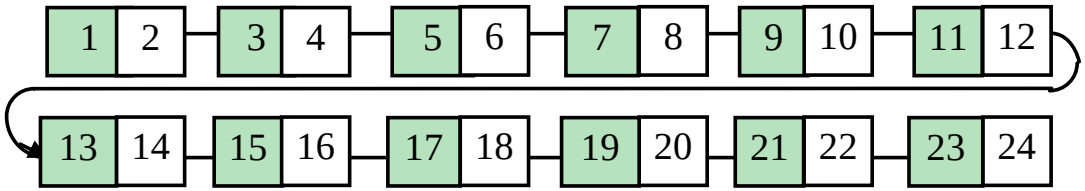
Half frame-rate front-end

Aurora 2 database, WER% [Tan et al., 2007]

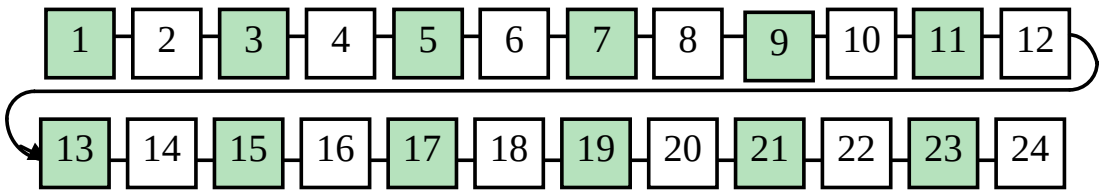
	16-state HMM	8-state HMM
Full frame rate	1.00	6.3
HFR-Duplication	1.02	5.84
HFR-NoDuplication	10.63	1.40

This motives a number of coding schemes (e.g. MDC, interleaving), which exploit temporal correlation of speech for error-robust and bandwidth-flexible DSR.

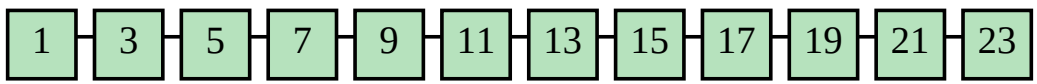
HFR motivated coding schemes



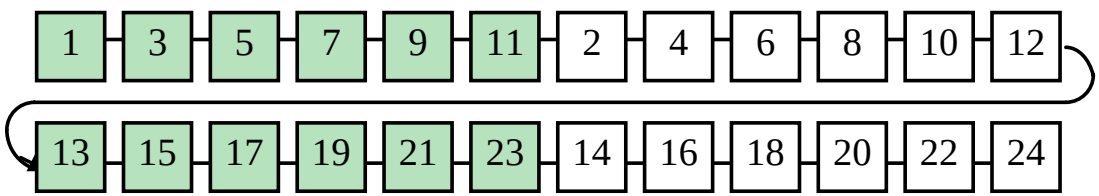
(a) ETSI-DSR front-end frame-pair scheme



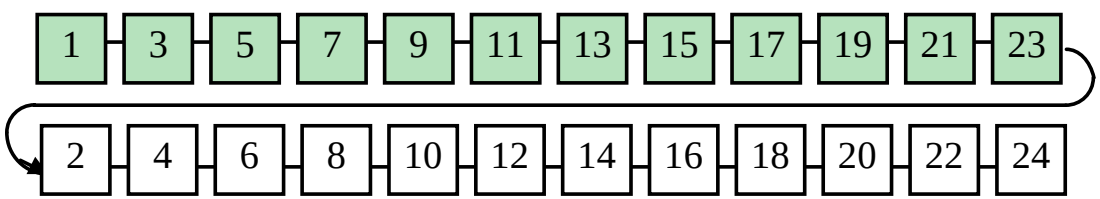
(b) One-frame scheme



(c) HFR scheme



(d) Interleaving12 scheme



(e) Interleaving24 scheme

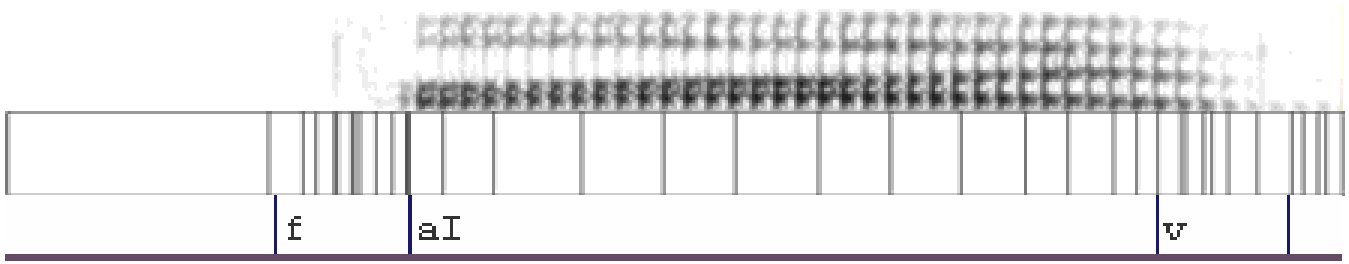
Variable frame-rate front-end

A posteriori SNR weighted energy based variable frame rate analysis [Tan and Lindberg, 2008]

- Frame selection based on the *a posteriori* SNR weighted energy distance of two consecutive frames:

$$D(t) = |\log E(t) - \log E(t-1)| \cdot SNR_{post}(t)$$

- Frame selection example



- Beneficial for source coding and noise robustness: at 1.5 kbps, WERs are 1.2% and 32.8% for clean and noisy speech (vs no compression: 1.0% and 38.7%).

Error-robustness performance on Aurora 2, EP3

	WER (%)	Bit-rate (bps)	Complexity	Compatibility with ETSI-DSR standards
Splicing	24.00	4 800	Low	Yes
No CRC	8.88	4 600	Low	No
Linear interpolation	7.35	4 800	Low	Yes
Repetition (Aurora)	6.70	4 800	Low	Yes
Weighted Viterbi	4.78	4 800	Low	Yes
RS(32, 16)	3.45	9 600	High	No
One-frame	3.41	5 000	Low	No
Uncertainty decoding	3.20	4800	Medium	Yes
Subvector	2.65	4 800	Low	Yes
Interleaving12	2.43	4 800	Low	No
Subvector + WVD	2.01	4 800	Low	Yes
Uncertainty decoding (inter-frame correlation)	1.98	4800	Medium	Yes
H-MAP	1.91	4 800	High	Yes
Interleaving24	1.74	4 800	Low	No
H-FBMMSE	1.34	4 800	High	Yes
MDC	1.04	5 200	Low	No
Error-free	0.95	4 800	-	-

1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

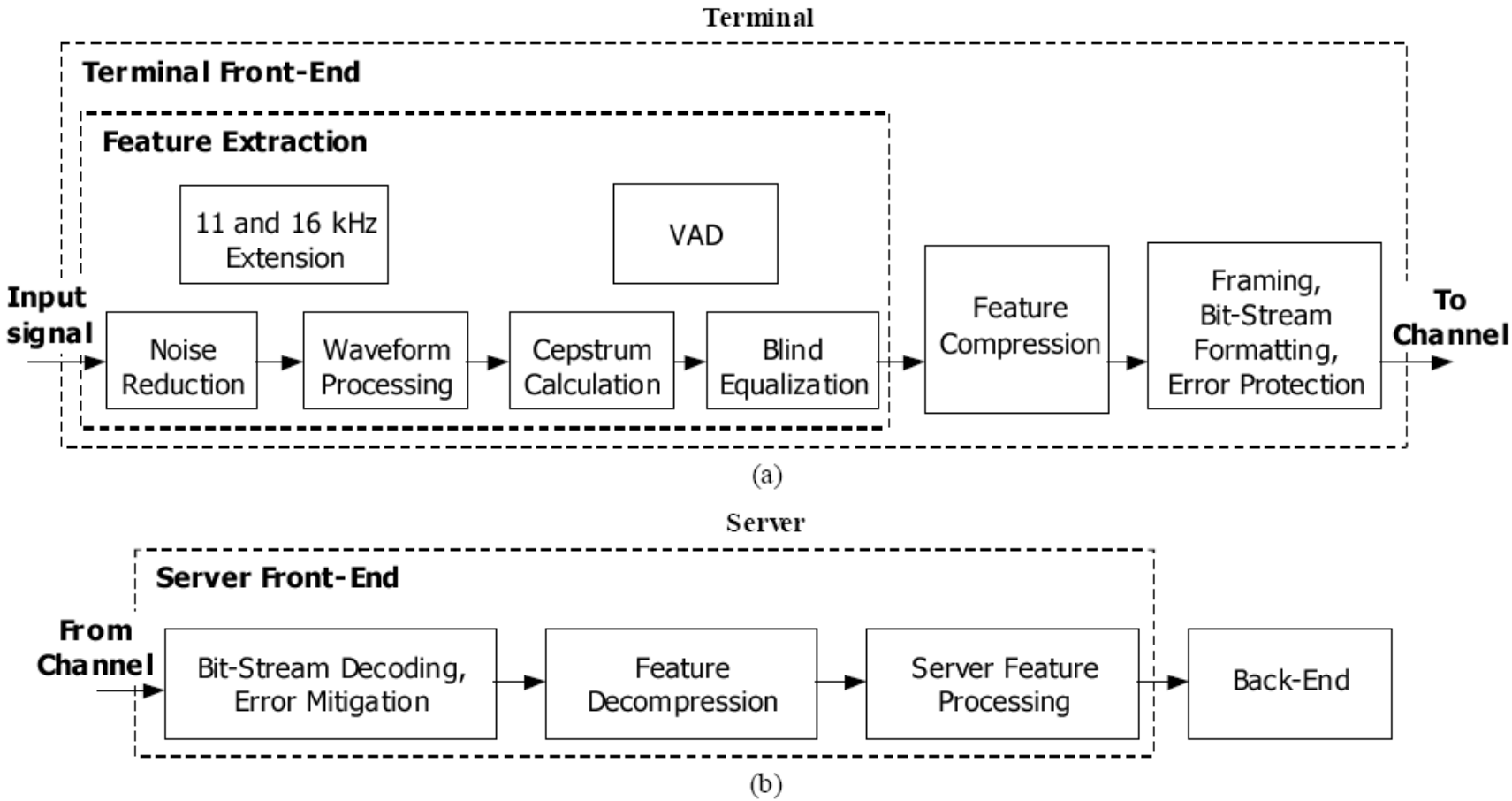
4 Embedded Speech Recognition

5 Applications

Overview of DSR standards

- Mel-cepstrum DSR front-end (FE) [ES 201 108]
  - ETSI STQ-Aurora, 2000
- Advanced DSR front-end (AFE) [ES 202 050]
  - ETSI STQ-Aurora, 2002
  - 53% error rate reduction in acoustic noise
- Extension for speech construction and tonal languages (XFE & XAFE) [ES 202 211], [ES 202 212]
  - ETSI STQ-Aurora, 2003
- Fixed point specifications for AFE and XAFE [3GPP TS 26.243]
  - 3GPP, 2004

Advanced front-end



From [ES 202 050]

Significant improvement over the basic front-end in noise robustness

Extended front-ends

Objectives of the extended front-ends

- Support speech construction and tonal languages.

Development trend of DSR and speech codecs:

- A convergence, though with different optimization objectives [Kim, 2008], [Milner and Shao, 2007].

AMR vs. DSR

Aurora databases (WER%) using AFE [Kelleher et al., 2002]

	DSR 4.4kbps	AMR 12.2kbps	AMR 4.75kbps
Aurora 2	12.6	15.3	18.7
Aurora 3	9.6	11.6	14.5

Aurora 2 database (WER%) [Kiss, 2000]

	EP1	EP2	EP3
GSM-EFR	3.02	4.35	12.87
DSR	2.01	2.06	8.98

Extensive comparison organised by 3GPP and conducted by industry [3GPP TR 26.943].

1 Introduction
- Devices and networks
- Automatic speech recognition

2 Network Speech Recognition
- Speech coding
- Transmission errors

3 Distributed Speech Recognition
- Properties of MFCCs
- Quantization
- Error recovery and concealment
- Standards
- Systems

4 Embedded Speech Recognition

5 Applications

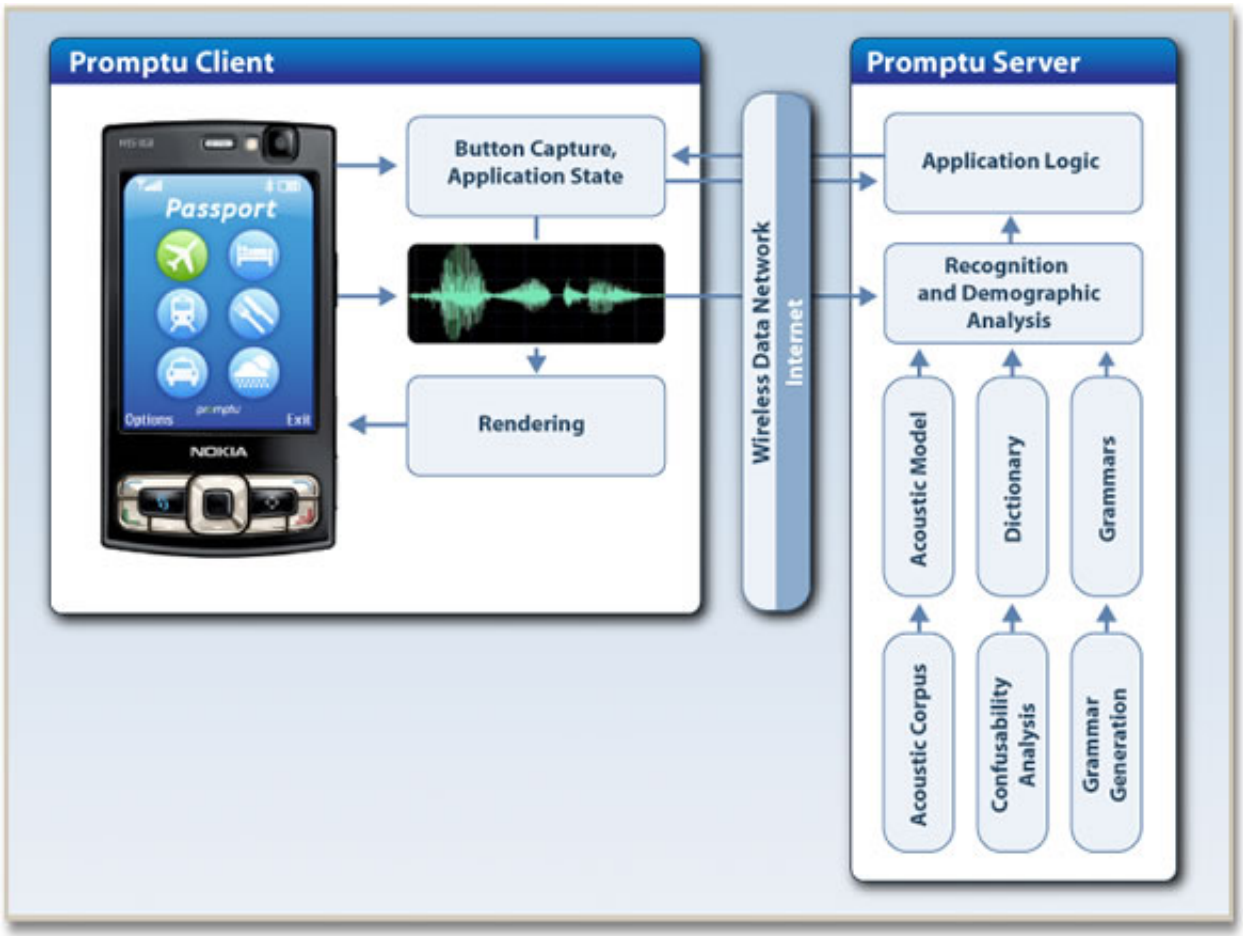
Remote speech recognition system

Microsoft®Response Point™is an innovative phone system software (VoIP enabled).

- "Response Point is an example of using the right technology for the right context and application. The blue button/voice recognition makes it easier for people to take the advantage of todays speech technology."
  - X.D. Huang

Remote speech recognition system

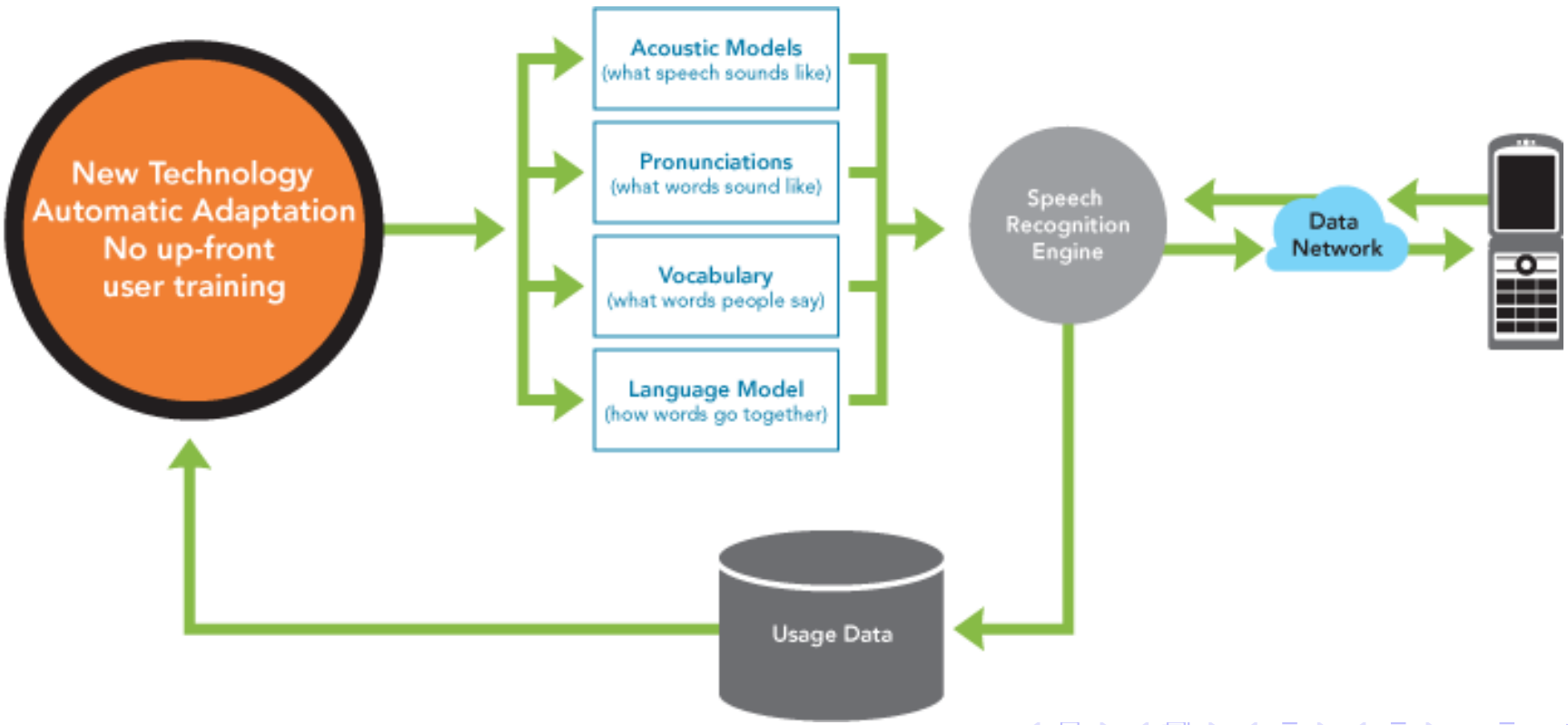
Promptu^{TM}provides multimodal solutions for mobile devices using client-server speech recognition technology.



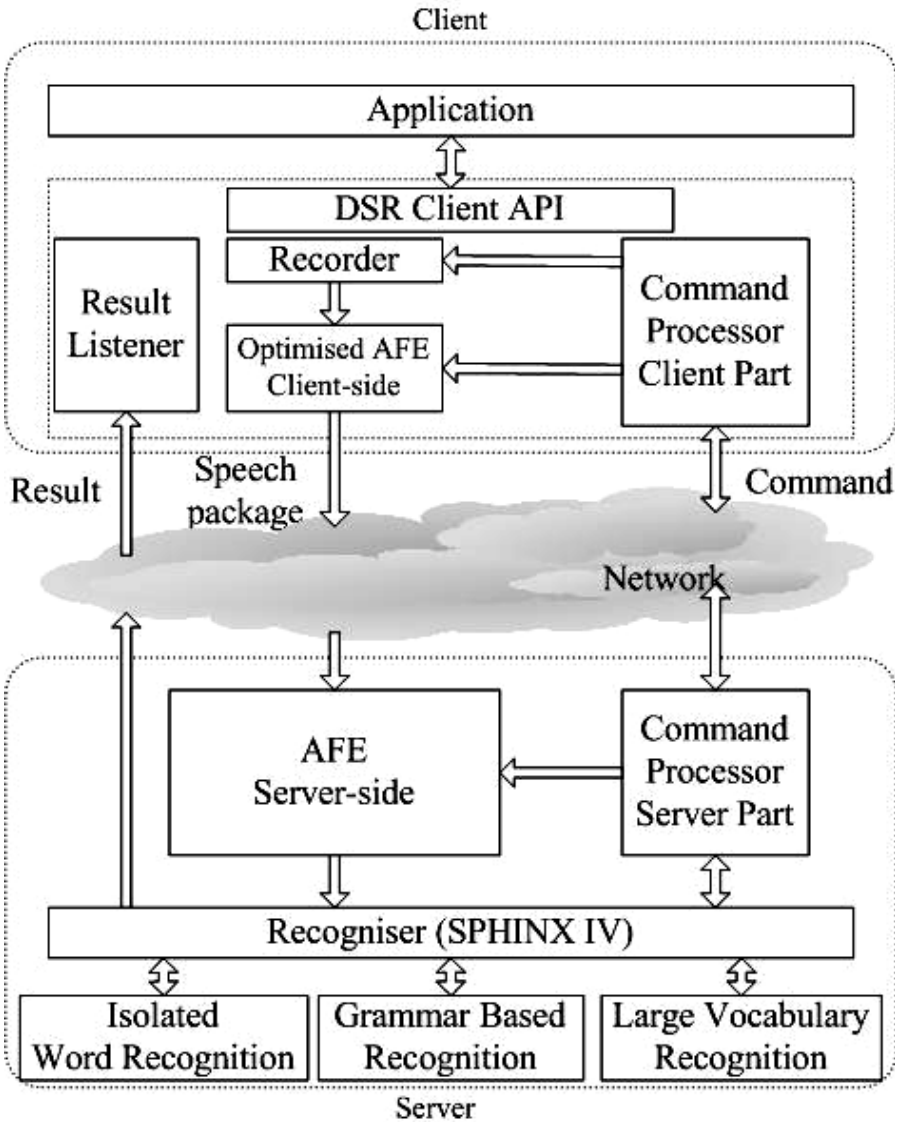
Remote speech recognition system

vlingo systems allow you to say anything to your mobile phone and still be recognized properly.

- Hierarchical Language Model Based Speech Recognition
- Adaptation



A configurable DSR system



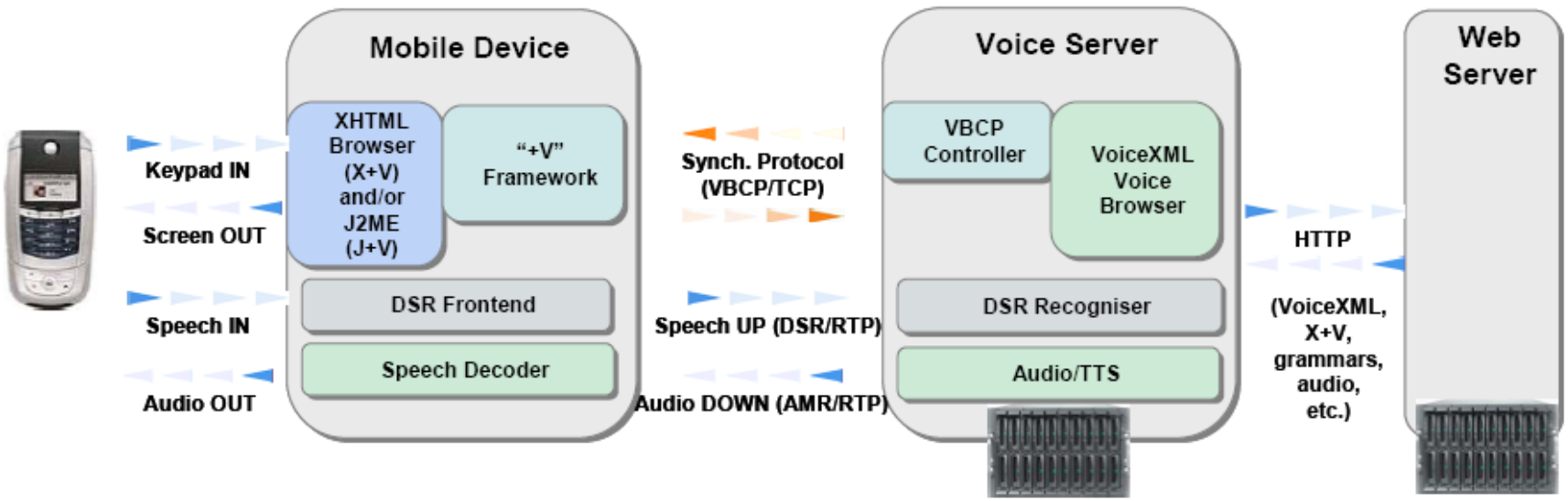
From [Xu et al., 2006].

A configurable DSR system

- Real-time efficiency in using different realisations of the AFE (Advance Front-End) and an H5550 IPAQ with a 400 MHz XScale CPU and 128 MB memory

Algorithm	FloatingP	FixedP	FixedP + FFT Optim.
X Real time	3.98	0.82	0.69

Distributed multimodal services



From [Pearce et al., 2005]

References and further reading I

Part I. Network and Distributed Speech Recognition

3GPP TS 26.243

"ANSI C Code for the fixed-point distributed speech recognition extended advanced front-end." 2004.

3GPP TR 26.943

"Recognition performance evaluations of codecs for Speech Enabled Services(SES)." 2004.

Besacier et al.

"The effect of speech and audio compression on speech recognition performance."
in *IEEE Multimedia Signal Processing Workshop*, Cannes, France, October 2001.

Borgstrom, B.J., Bernard, A. and Alwan, A.

"Error recovery: channel coding and packetization."
in *Z.-H. Tan, and B. Lindberg (eds.), Automatic speech recognition on mobile devices and over communication networks, Springer*, 2008.

Bryant, R.

"Data-intensive supercomputing: The case for DISC."
CMU Technical Report CMU-CS-07-128, May 2007.

Cardenal-Lopez et al.

"Soft decoding strategies for distributed speech recognition over IP newtworks."
in *Proc. ICASSP*, Montreal, Canada, 2004.

Digalakis, V., Neumeyer, L. and Perakakis, M.

"Quantization of cepstral parameters for speech recognition over the World Wide Web."
IEEE J. Select. Areas Communications, vol. 17, no. 1, pp. 82-90, 1999.

References and further reading II

ETSI Standard ES 201 108

"Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithm, v1.1.2." 2000.

ETSI Standard ES 202 050

"Distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithm." 2002.

ETSI Standard ES 202 211

"Distributed speech recognition; extended front-end feature extraction algorithm; compression algorithm, back-end speech reconstruction algorithm." 2003.

ETSI Standard ES 202 212

"Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm, back-end speech reconstruction algorithm." 2003.

Gomez, A.M., Peinado, A.M., Sanchez, V., and Rubio, A.J.

"A source model mitigation technique for distributed speech recognition over lossy packet channels."
in *Proc. Eurospeech*, Geneva, Switzerland, 2003.

Hirsch, H.G.

"The influence of speech coding on recognition performance in telecommunication networks."
in *Proc. ICSLP*, Denver, USA, September 2002.

Hirsch, H.G. and Pearce D.

"The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions."
in *Proc. ISCA ITRW ASR*, Paris, France, 2000.

References and further reading III

Ho, Y.-C.

"The no free lunch theorem and the human-machine interface."
IEEE Control Syst., 8-10, June 1999.

Ion, V. and Haeb-Umbach, R.

"Uncertainty decoding for distributed speech recognition over error-prone networks,"
Speech Communication, vol. 48, pp. 1435-1446, 2006.

James, A.B. and Milner, B.P.

"Towards improving the robustness of distributed speech recognition in packet loss."
in *Proc. COST278 & ISCA Research Workshop on Robustness Issues in Conversational Interaction*, Norwich, UK, 2004.

Kelleher, H, Pearce, D., Ealey, D. and Mauuary, L.

"Speech recognition performance comparison between DSR and AMR transcoded speech."
in *Proc. ICSLP*, Denver, USA, September 2002.

Kim, H.K.

"Speech recognition over IP networks"
in *Z.-H. Tan, and B. Lindberg (eds.), Automatic speech recognition on mobile devices and over communication networks, Springer*, 2008.

Kim, H.K. and Cox, R.V.

"A bitstream-based front-end for wireless speech recognition on IS-136 communications system."
IEEE Trans. Speech and Audio Processing, vol. 9, no. 5, pp. 558-568, 2001.

Kiss, I.

"A comparison of distributed and network speech recognition for mobile communication systems."
in *Proc. ICSLP*, Beijing, China, October 2000.

References and further reading IV

Milner, B. P. and James, A. B.

"Robust speech recognition over mobile and IP networks in burst-like packet loss."
IEEE Trans. Audio, Speech and Language Processing, vol. 14, no. 1, pp. 223-231, 2006.

Milner, B. and Shao, X.

"Prediction of fundamental frequency and voicing from Mel-frequency cepstral coefficients for unconstrained speech reconstruction."
IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 1, pp. 24-33, 2007.

Pearce, D., Engelsma, J., Ferrans, J. and Johnson, J.

"An architecture for seamless access to distributed multimodal services."
in *Proc.INTERSPEECH*, Lisbon, Portugal, September 2005.

Peinado, A., Sanchez, V., Perez-Cordoba, J., and de la Torre, A.

"HMM-based channel error mitigation and its application to distributed speech recognition."
Speech Communication, vol. 41, pp. 549-561, 2003.

S.So, and K.K. Paliwal,

"Scalable distributed speech recognition using Gaussian mixture model-based block quantization,"
Speech Communication, vol. 48, pp. 746–758, 2006.

S.So, and K.K. Paliwal,

"Quantization of speech features: Source coding."
in *Z.-H. Tan, and B. Lindberg (eds.), Automatic speech recognition on mobile devices and over communication networks, Springer*, 2008.

Tan, Z.-H., Dalsgaard, P. and Lindberg, B.

"Automatic speech recognition over error-prone wireless networks."
Speech Communication, vol. 47, no. 1-2, pp. 220-242, 2005.

References and further reading V

Tan, Z.-H., Dalsgaard, P. and Lindberg, B.

"Exploiting temporal correlation of speech for error-robust and bandwidth-flexible distributed speech recognition."
IEEE Trans. on Audio, Speech and Language Processing, vol. 15, no. 4, pp. 1391-1403, 2007.

Tan, Z.-H. and Lindberg, B.

"A Posteriori SNR Weighted Energy Based Variable Frame Rate Analysis for Speech Recognition."
in *Proc. Interspeech*, Brisbane, Australia, 2008.

Z.-H..Tan, and I. Varga,

"Network, distributed and embedded speech recognition: an overview,"
in *Z.-H. Tan, and B. Lindberg (eds.), Automatic speech recognition on mobile devices and over communication networks, Springer*, 2008.

Wan, C.-Y. and Lee, L.-S.

"Histogram-based quantization for robust and/or distributed Speech Recognition."
IEEE Trans. on Audio, Speech and Language Processing, vol. 16, no. 4, pp. 859-873, 2008.

Xu, H., Tan, Z.-H., Dalsgaard, P., Mattethat, R. and Lindberg, B.

"A configurable distributed speech recognition system."
in *H. Abut, J.H.L. Hansen, K. Takeda (eds.), Digital Signal Processing for In-Vehicle and Mobile Systems 2*, Springer Science, New York, 2006.

Zhu, Q. and Alwan, A.

"An efficient and scalable 2D DCT-based feature coding scheme for remote speech recognition."
in *Proc. ICASSP*, Salt Lake City, USA, 2001.