MPEG-4 TTS 와 음성통역 아바타

양 재 우

Human Interface Technology Department

ETRI-STTL

VOL.36, NO.8, Aug., 1988.

[17] "Principles of Vector-Sum Excited Linear Predictive (VSELP) Speech Coder and its Implementation on the DSP56156," Motorola, Inc., Digital Processor Operations, Austin, Texas.

[18] I.A.Gerson and M.A.Jasiuk, "A 5600 bps VSELP Speech Coder Candidate for Half rate GSM," EUROSPEECH'93, 1993.

[19] R.Salami, C.Laflamme and J.P.Adoul, "8kbit/s ACELP coding of speech with 10ms speech frame: A candidate for CCITT standardization", ICASSP'94, pp. II-97-II-100.

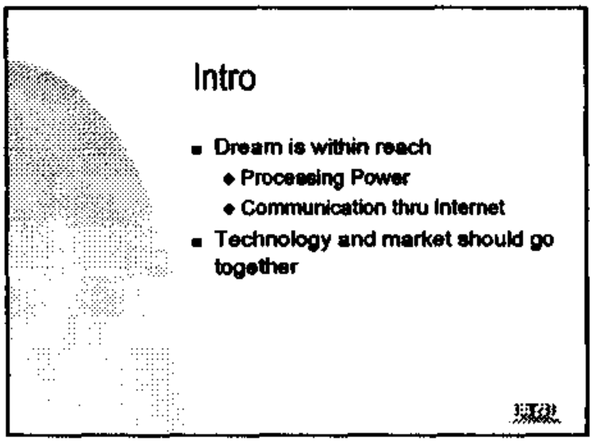
[20] Akitoshi Kataoka, Takehiro Moriya and Shinji Hayashi, "Implementation and Performance of an 8-kbit/s Conjugate Structure CELP Speech Coder", ICASSP'94., pp.II-93-96.

[21] Akitoshi Kataoka, Takehiro Moriya and Shinji Hayashi, "An 8-kbit/s Speech Coder Based on Conjugate Structure CELP", ICASSP'93, pp. II-592-595.

[22] 김 홍국, 김 삼룡, "PCS를 위한 음성코딩방식과 음질비교", 전자공학회지 22권 9호, pp.1060 -1066., 1995.

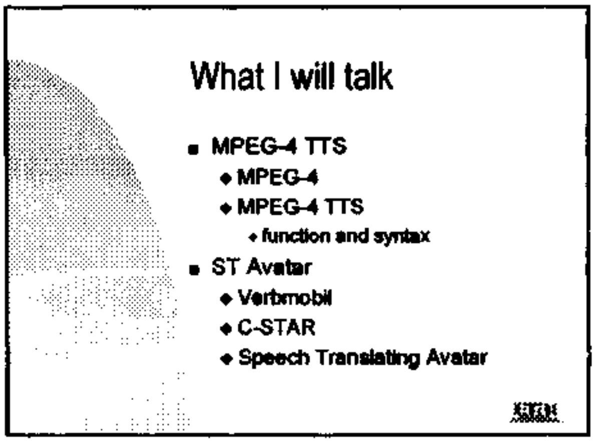
[23] T. Ohya, H. Suda and T.Miki, "5.6 kbits/s PSI-CELP of the Half-rate PDC Speech Coding Standard", IEEE VTC, pp.1680-1684, 1994.

[24] TIA/EIA/IS-127 Enhanced Variable Rate Codec, Speech Service Option 3 for Wideband Spread Spectrum Digital Systems.



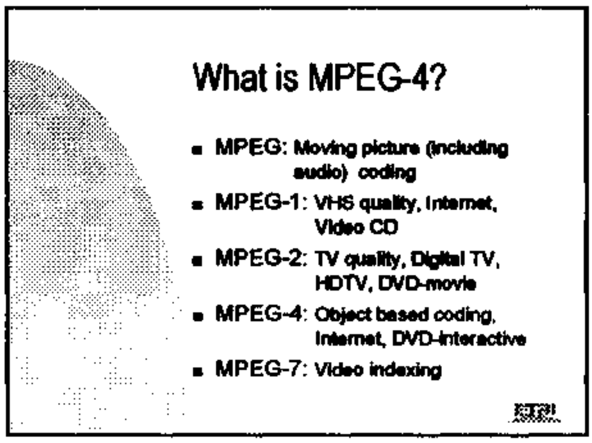
Intro

- Dream is within reach
  - ◆ Processing Power
  - ◆ Communication thru Internet
- Technology and market should go together



What I will talk

- MPEG-4 TTS
  - ◆ MPEG-4
  - ◆ MPEG-4 TTS
    - ◆ function and syntax
- ST Avatar
  - ◆ Verbmobil
  - ◆ C-STAR
  - ◆ Speech Translating Avatar



What is MPEG-4?

- MPEG: Moving picture (including audio) coding
- MPEG-1: VHS quality, Internet, Video CD
- MPEG-2: TV quality, Digital TV, HDTV, DVD-movie
- MPEG-4: Object based coding, Internet, DVD-interactive
- MPEG-7: Video indexing

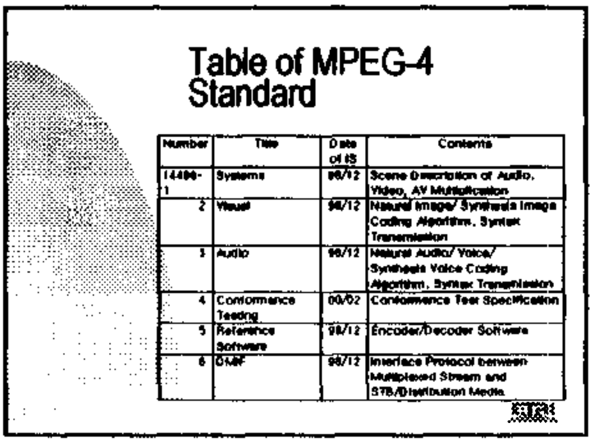
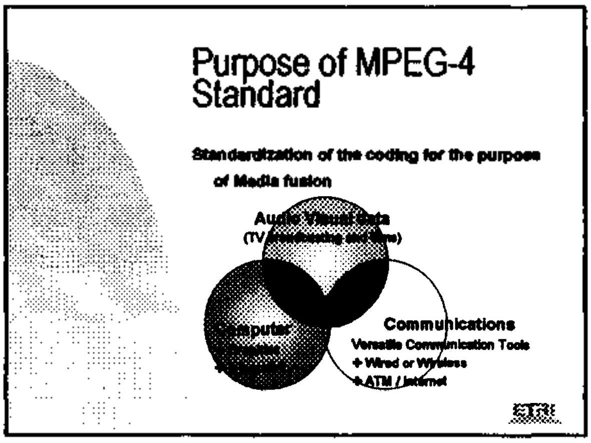


Table of MPEG-4 Standard

Number	Title	Date of IS	Contents
14496-1	Systems	98/12	Scene Description of Audio, Video, AV Multiplication
2	Visual	98/12	Natural image/ Synthesis image Coding Algorithm, Syntax Transmission
3	Audio	98/12	Natural Audio/ Voice/ Synthesis Voice Coding Algorithm, Syntax Transmission
4	Conformance Testing	00/02	Conformance Test Specification
5	Reference Software	98/12	Encoder/Decoder Software
6	DMIF	98/12	Interface Protocol between Multiplexed Stream and STB/Distribution Media



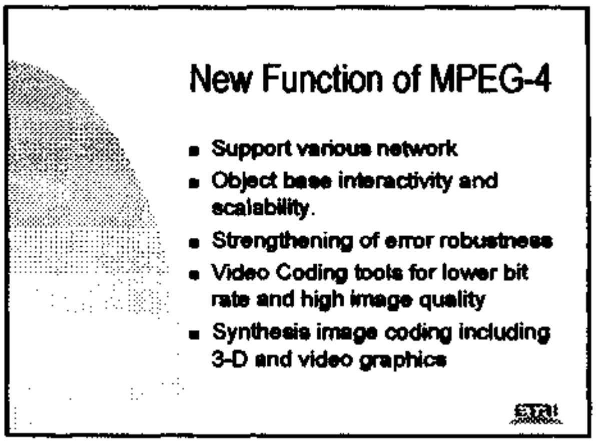
Purpose of MPEG-4 Standard

Standardization of the coding for the purpose of Media fusion

Audio Visual Data
(TV [illegible])

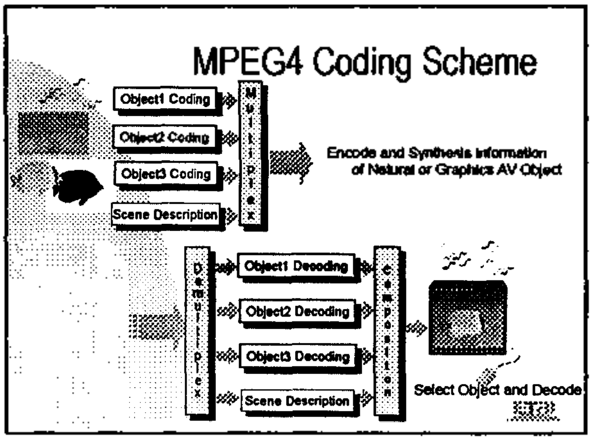
Computer
[illegible]

Communications
Versatile Communication Tools
◆ Wired or Wireless
◆ ATM / Internet



New Function of MPEG-4

- Support various network
- Object base interactivity and scalability.
- Strengthening of error robustness
- Video Coding tools for lower bit rate and high image quality
- Synthesis image coding including 3-D and video graphics



MPEG4 Coding Scheme

Object1 Coding
Object2 Coding
Object3 Coding
Scene Description

Multiplex

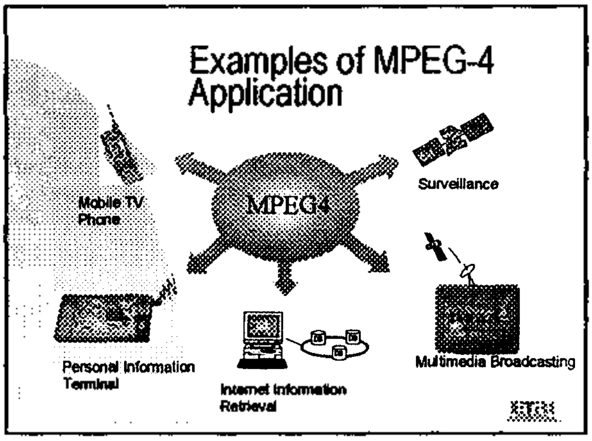
Encode and Synthesis Information of Natural or Graphics AV Object

Demultiplex

Object1 Decoding
Object2 Decoding
Object3 Decoding
Scene Description

Composition

Select Object and Decode



Examples of MPEG-4 Application

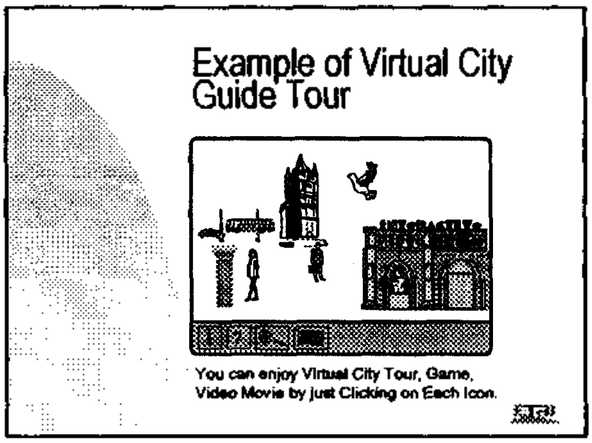
Mobile TV Phone

Surveillance

Personal Information Terminal

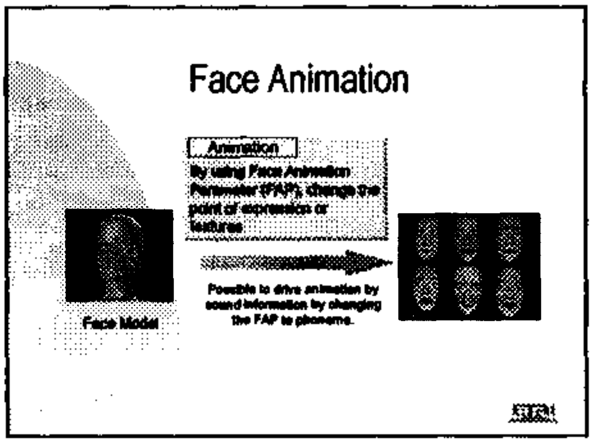
Internet Information Retrieval

Multimedia Broadcasting



Example of Virtual City Guide Tour

You can enjoy Virtual City Tour, Game, Video Movie by just Clicking on Each Icon.



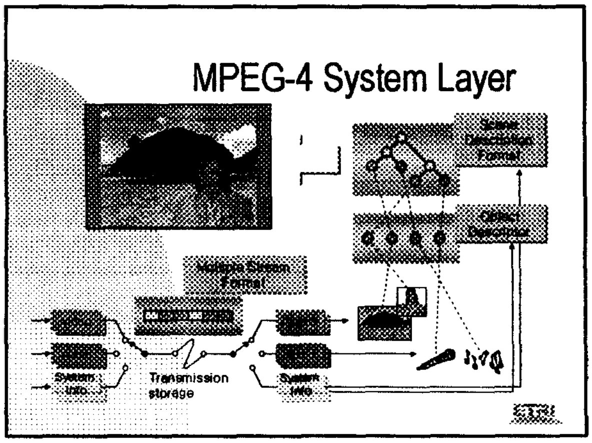
Face Animation

Animation

By using Face Animation Parameter (FAP), change the point of expression or features

Face Model

Possible to drive animation by sound information by changing the FAP to phoneme.



MPEG-4 System Layer

Scene Description Format

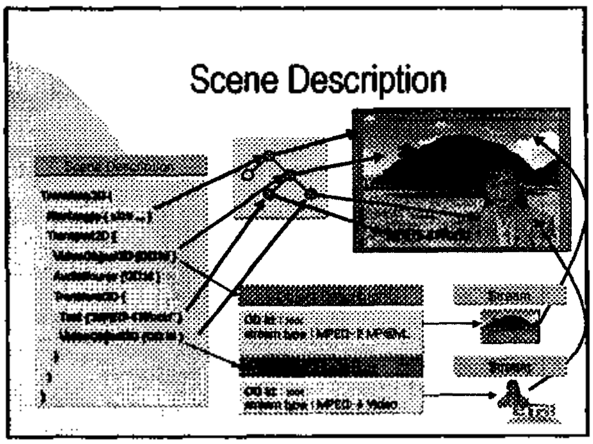
Object Descriptor

Multiple Stream Format

System Info

Transmission storage

System Info



Scene Description

Scene Description

Transform2D {
Rectangle (size ...)
Transform2D {
VideoObject2D (OD30)
AudioSource (OD34)
Transform2D {
Text ("MPEG-4 Model")
VideoObject2D (OD 8)
}
}
}

[illegible]

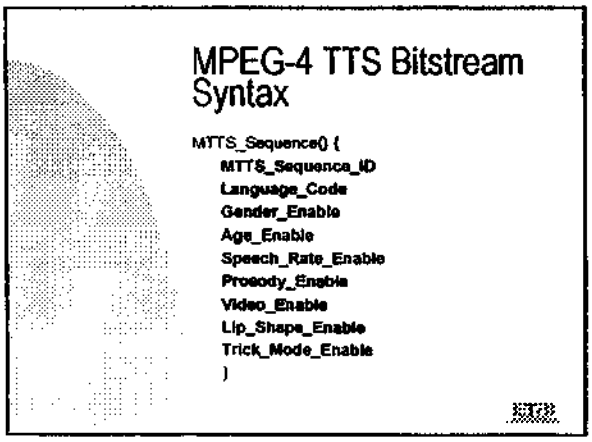
OD Id : xxx
stream type : MPEG-2 MP@ML

[illegible]

OD Id : xxx
stream type : MPEG-4 Video

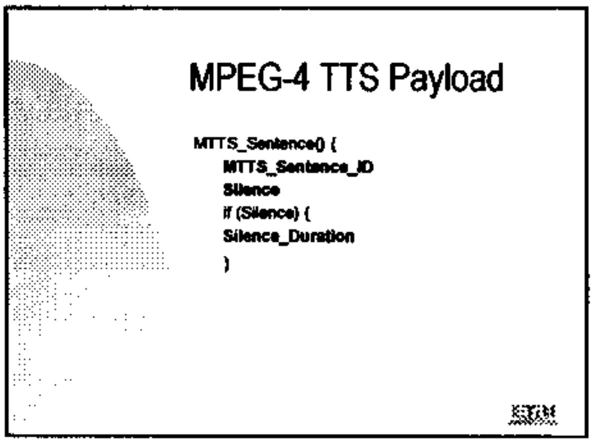
Stream

Stream



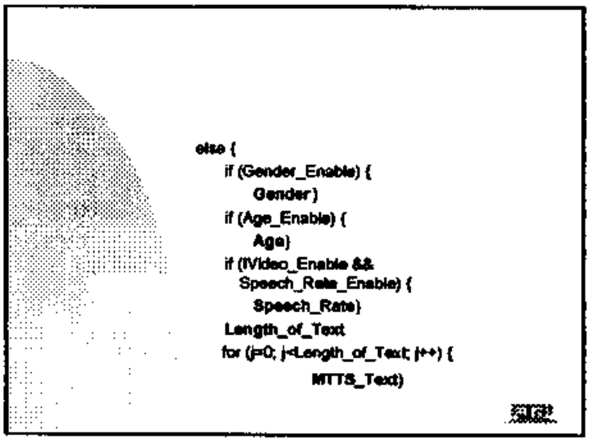
MPEG-4 TTS Bitstream Syntax

```
MTTS_Sequence() {
    MTTS_Sequence_ID
    Language_Code
    Gender_Enable
    Age_Enable
    Speech_Rate_Enable
    Prosody_Enable
    Video_Enable
    Lip_Shape_Enable
    Trick_Mode_Enable
    }
```

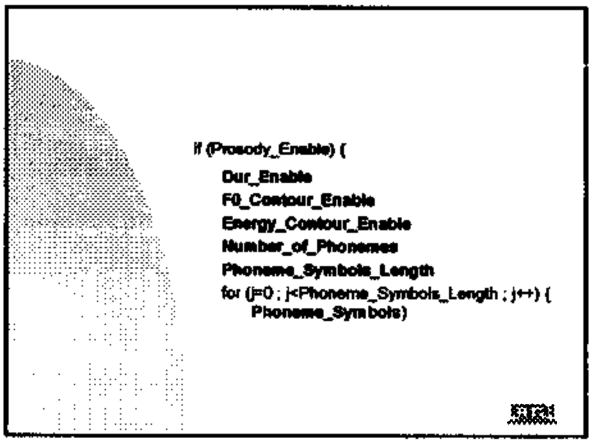


MPEG-4 TTS Payload

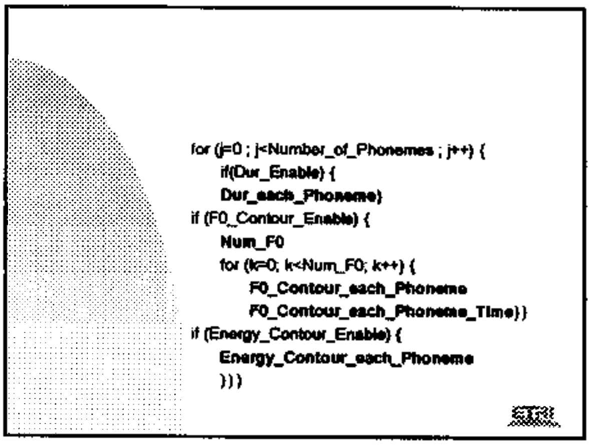
```
MTTS_Sentence() {
    MTTS_Sentence_ID
    Silence
    if (Silence) {
    Silence_Duration
    }
```



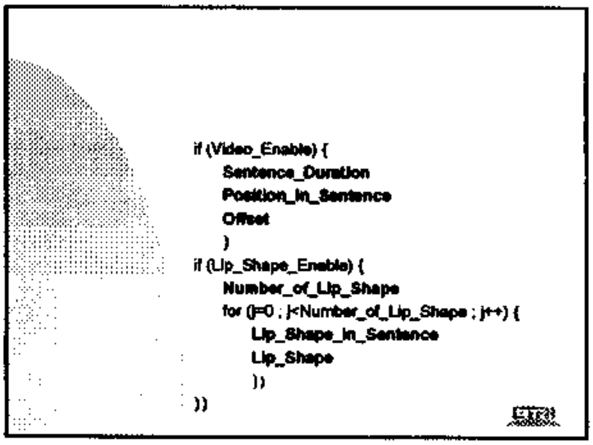
```
else {
    if (Gender_Enable) {
        Gender}
    if (Age_Enable) {
        Age}
    if (!Video_Enable &&
      Speech_Rate_Enable) {
        Speech_Rate}
    Length_of_Text
    for (j=0; j<Length_of_Text; j++) {
                MTTS_Text}
```



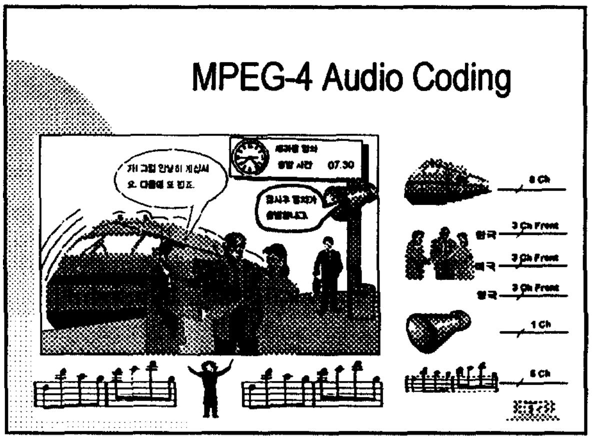
```
if (Prosody_Enable) {
    Dur_Enable
    F0_Contour_Enable
    Energy_Contour_Enable
    Number_of_Phonemes
    Phoneme_Symbols_Length
    for (j=0 ; j<Phoneme_Symbols_Length ; j++) {
        Phoneme_Symbols}
```



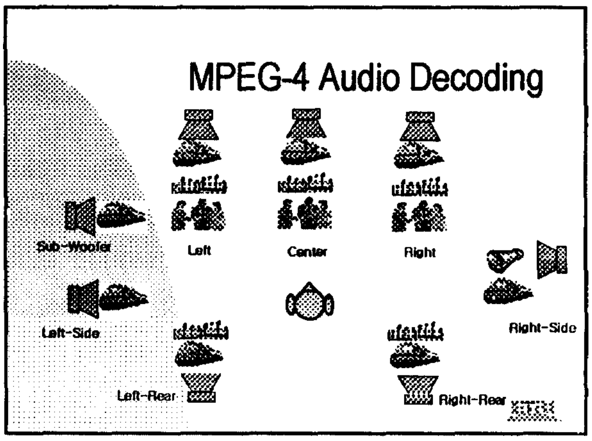
```
for (j=0 ; j<Number_of_Phonemes ; j++) {
    if(Dur_Enable) {
    Dur_each_Phoneme}
if (F0_Contour_Enable) {
    Num_F0
    for (k=0; k<Num_F0; k++) {
        F0_Contour_each_Phoneme
        F0_Contour_each_Phoneme_Time}}
if (Energy_Contour_Enable) {
    Energy_Contour_each_Phoneme
    }}}
```



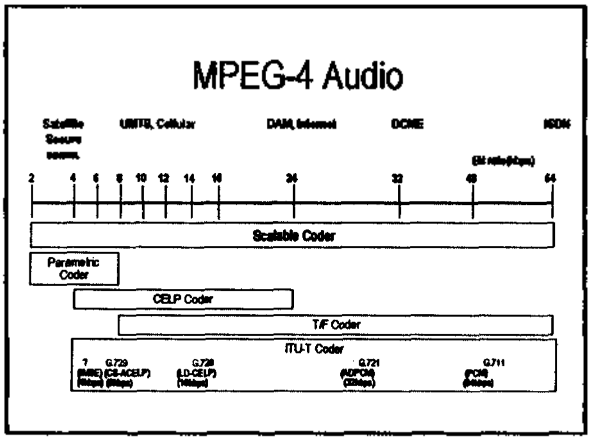
```
if (Video_Enable) {
    Sentence_Duration
    Position_in_Sentence
    Offset
    }
if (Lip_Shape_Enable) {
    Number_of_Lip_Shape
    for (j=0 ; j<Number_of_Lip_Shape ; j++) {
        Lip_Shape_in_Sentence
        Lip_Shape
        }}
}}
```



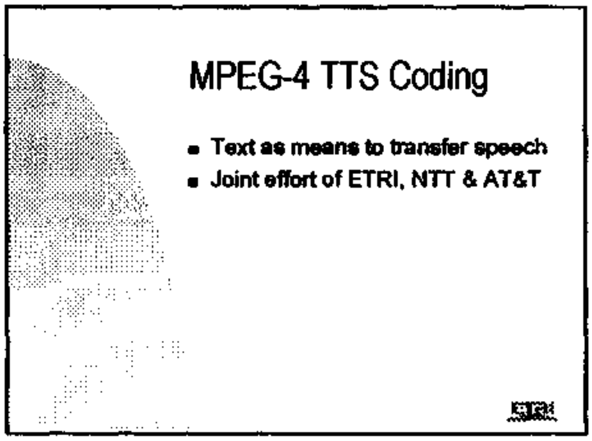
MPEG-4 Audio Coding



MPEG-4 Audio Decoding

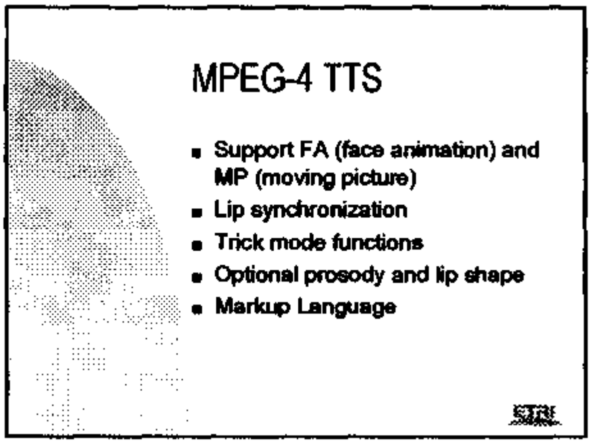


MPEG-4 Audio



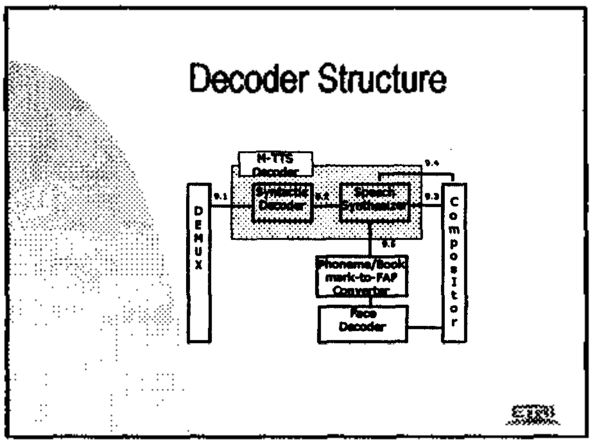
MPEG-4 TTS Coding

- Text as means to transfer speech
- Joint effort of ETRI, NTT & AT&T

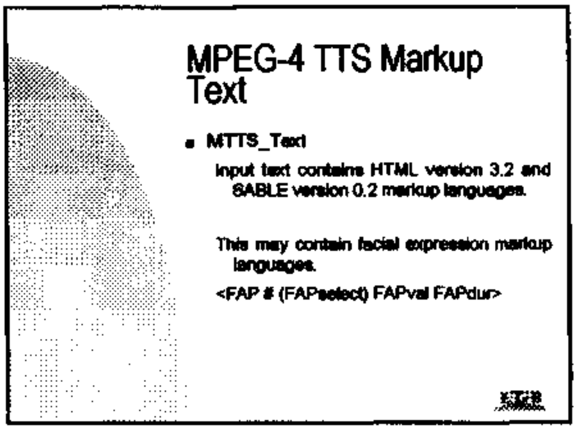


MPEG-4 TTS

- Support FA (face animation) and MP (moving picture)
- Lip synchronization
- Trick mode functions
- Optional prosody and lip shape
- Markup Language



Decoder Structure



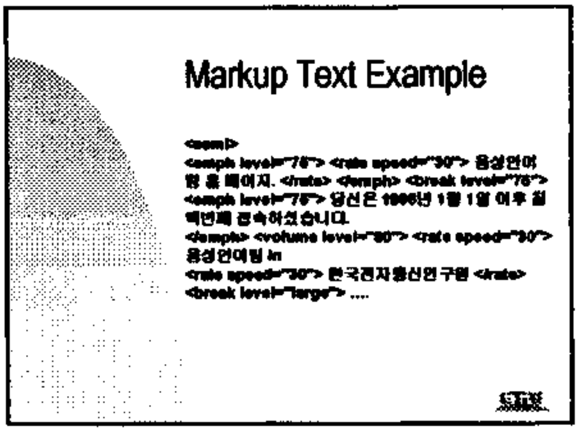
MPEG-4 TTS Markup Text

- MTTS_Text

  Input text contains HTML version 3.2 and SABLE version 0.2 markup languages.

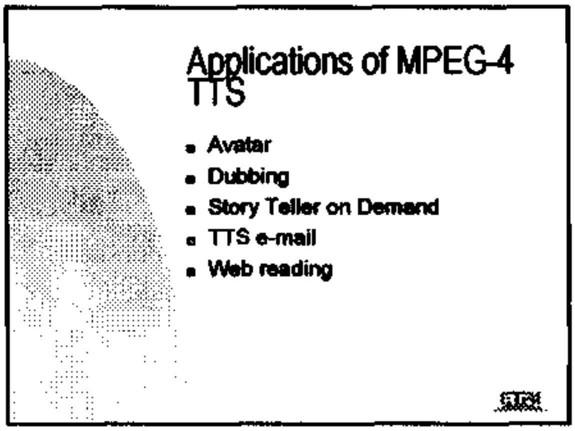
  This may contain facial expression markup languages.

  <FAP # (FAPselect) FAPval FAPdur>



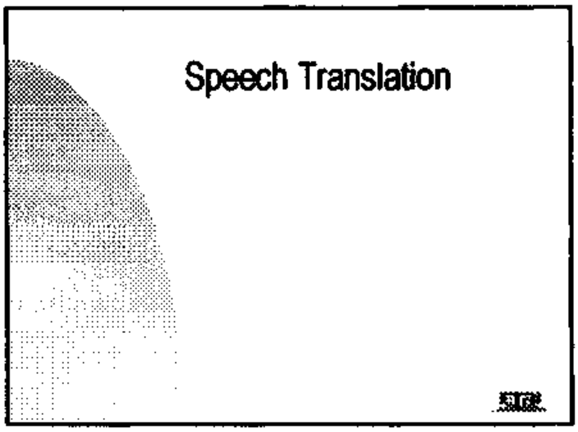
Markup Text Example

<sem/>
<emph level="75"> <rate speed="30"> 음성언어
방 홈 페이지. </rate> </emph> <break level="75">
<emph level="75"> 당신은 1996년 1월 1일 이후 일
백번째 접속하셨습니다.
</emph> <volume level="80"> <rate speed="30">
음성언어방 in
<rate speed="30"> 한국전자통신연구원 </rate>
<break level="large"> ....

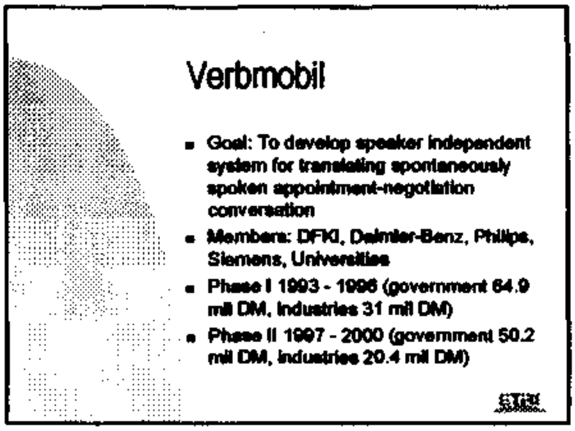


Applications of MPEG-4 TTS

- Avatar
- Dubbing
- Story Teller on Demand
- TTS e-mail
- Web reading

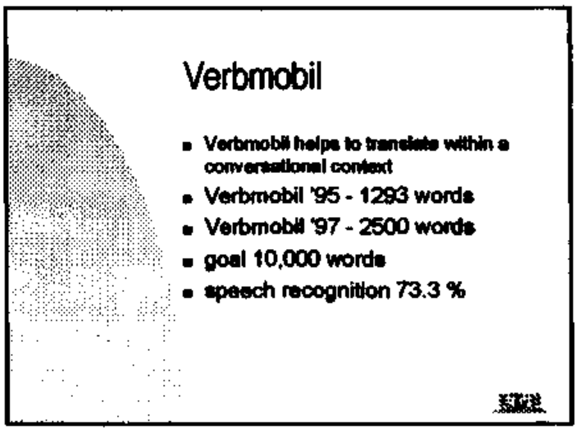


Speech Translation



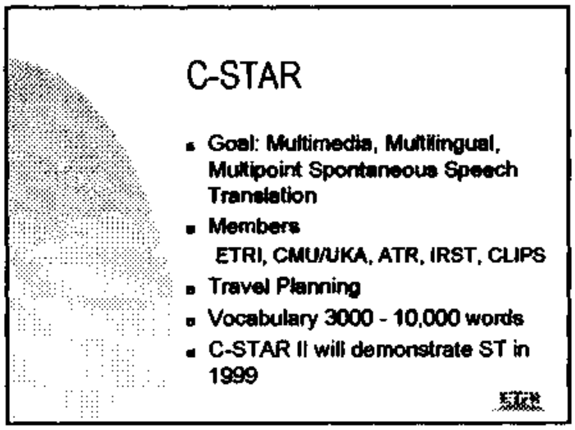
Verbmobil

- Goal: To develop speaker independent system for translating spontaneously spoken appointment-negotiation conversation
- Members: DFKI, Daimler-Benz, Philips, Siemens, Universities
- Phase I 1993 - 1996 (government 64.9 mil DM, industries 31 mil DM)
- Phase II 1997 - 2000 (government 50.2 mil DM, industries 20.4 mil DM)



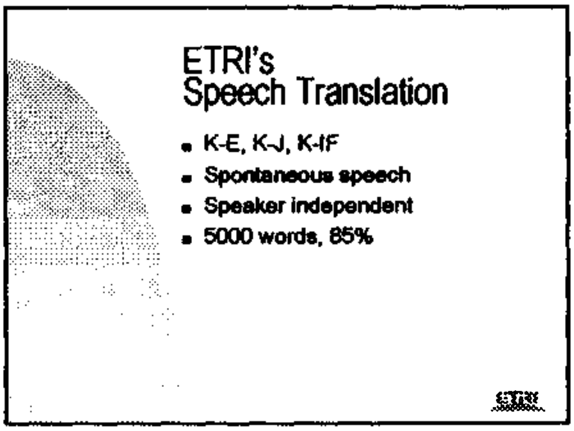
Verbmobil

- Verbmobil helps to translate within a conversational context
- Verbmobil '95 - 1293 words
- Verbmobil '97 - 2500 words
- goal 10,000 words
- speech recognition 73.3 %



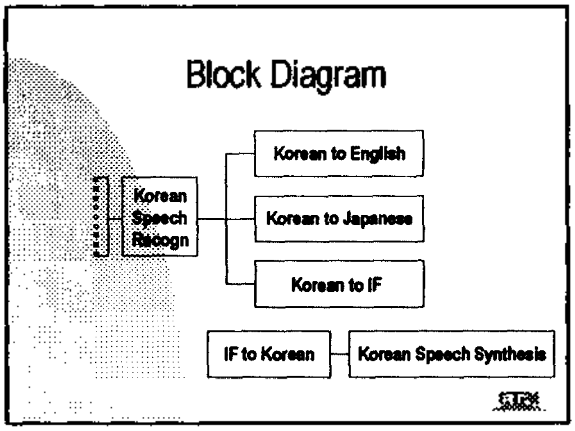
C-STAR

- Goal: Multimedia, Multilingual, Multipoint Spontaneous Speech Translation
- Members
  ETRI, CMU/UKA, ATR, IRST, CLIPS
- Travel Planning
- Vocabulary 3000 - 10,000 words
- C-STAR II will demonstrate ST in 1999

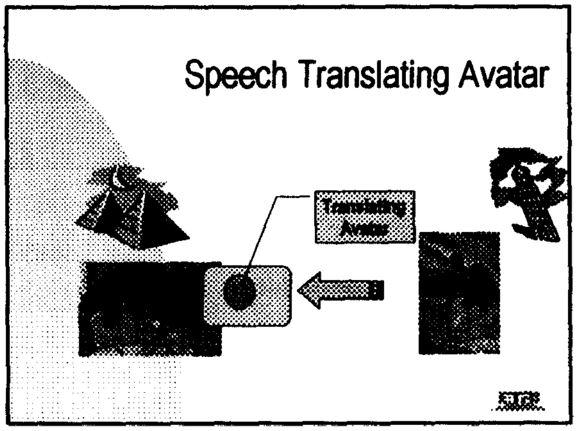


ETRI's Speech Translation

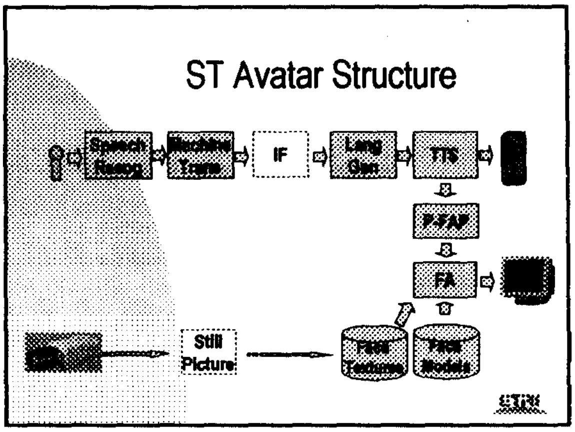
- K-E, K-J, K-IF
- Spontaneous speech
- Speaker independent
- 5000 words, 85%



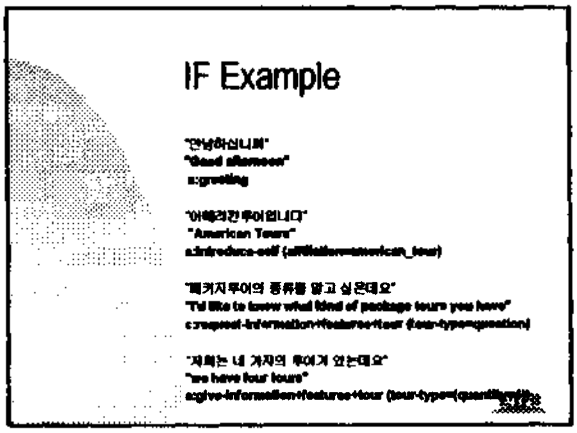
Block Diagram



Speech Translating Avatar



ST Avatar Structure



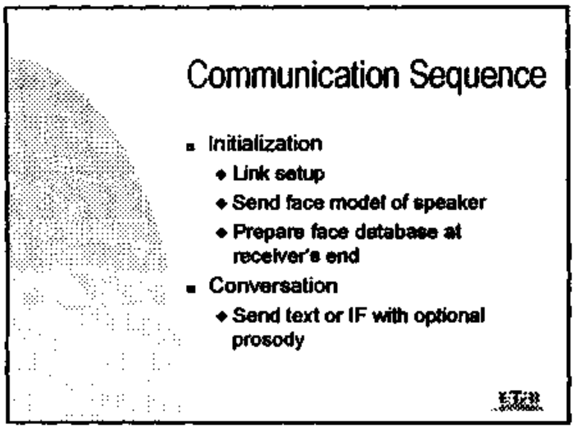
IF Example

"안녕하십니까"
"Good afternoon"
a:greeting

"아메리칸 투어입니다"
"American Tours"
a:introduce-self (affiliation=american_tour)

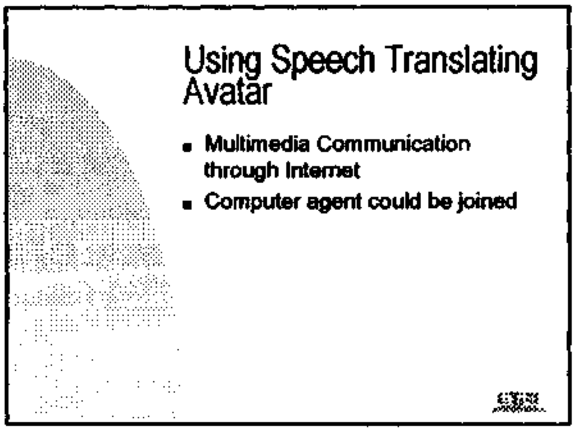
"패키지투어의 종류를 알고 싶은데요"
"I'd like to know what kind of package tours you have"
c:request-information+features+tour (tour-type=question)

"저희는 네 가지의 투어가 있는데요"
"we have four tours"
a:give-information+features+tour (tour-type=(quantity=4



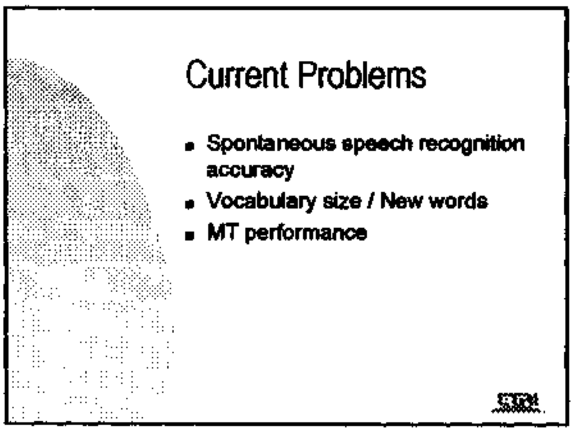
Communication Sequence

- Initialization
  - Link setup
  - Send face model of speaker
  - Prepare face database at receiver's end
- Conversation
  - Send text or IF with optional prosody



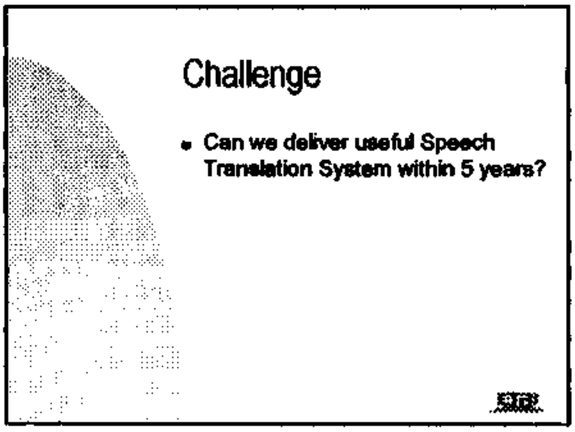
Using Speech Translating Avatar

- Multimedia Communication through Internet
- Computer agent could be joined



Current Problems

- Spontaneous speech recognition accuracy
- Vocabulary size / New words
- MT performance



Challenge

- Can we deliver useful Speech Translation System within 5 years?