



DP - Dynamisk programmering

snabb, optimal kombinerad av delbeslut

Analogi

- Sök bästa väg mellan två adresser i en stad.
- Ett mycket stort antal möjliga vägar, men: vid varje korsning är den bästa fortsättningen oberoende av hur man kom till denna punkt. Endast den bästa vägen hit kan alltså ingå i den totalt bästa vägen. Övriga vägar hit innebär en sämre totalväg och strykes.
- Antalet sökningar minskar dramatiskt.



GSLT Tal- och talatigenkänning M. Blomberg, 2002-9-9 | 31 |



Exempel på DP - textjämförelse

Jämförelse mellan textsträngar

Sträng A: "ALDRIG" (korrekt, referens)

Sträng B: "ALLDRIG" (okänt)

Frågeställning: Vad är den minimala "distansten" mellan strängarna efter bästa synkronisering?

(Avsåg författaren troligen detta ord? Är det felstavat? I så fall hur?)

Lokal distans:

$d[i,j] = 0$ om $A[i] = B[j]$; annars $= 1$.

Global (ackumulerad) distans:

$D[i,j] = \min(D[i-1,j], D[i-1,j-1], D[i,j-1]) + d[i,j]$

Initiering:

$D[0,0] = d[0,0]$

Förenklad algoritm:

for($i=1$; $i \leq I$; $i++$)

for($j=1$; $j \leq J$; $j++$)

$D[i,j] = d[i,j] + \min(D[i-1,j], D[i-1,j-1], D[i,j-1]);$

return($D[I,J]$);

GSLT Tal- och talatigenkänning M. Blomberg, 2002-9-9 | 32 |



DP-exempel (forts.)

Lokala distanser $d[i,j]$

G	1	1	1	1	1	0
I	1	1	1	1	1	0
R	1	1	1	1	0	1
D	1	1	1	0	1	1
L	1	0	0	1	1	1
A	0	1	1	1	1	1
A	L	L	D	R	I	G

Förenklad algoritm:

for($i=1$; $i \leq I$; $i++$)
for($j=1$; $j \leq J$; $j++$)
 $D[i,j] = d[i,j] + \min(D[i-1,j], D[i-1,j-1], D[i,j-1]);$
return($D[I,J]$);

Ackumulerade distanser $D[i,j]$

G	5	4	4	3	2	1	0
I	4	3	3	2	1	0	1
R	3	2	2	1	0	1	2
D	2	1	1	0	1	2	3
L	1	0	0	1	2	3	4
A	0	1	2	3	4	5	6
A	L	L	D	R	I	G	

Textdistans med DP: 0
(Distans tecken-för-tecken: 5)

GSLT Tal- och talatigenkänning M. Blomberg, 2002-9-9 | 33 |



För- och nackdelar med ordmönsterigenkänning

+ Kräver ej fonetisk beskrivning

+ Koartikulation inom ord tas om hand

- Koartikulation i ordgräns tas ej om hand

- Kräver inläsning av alla ord vid träning

- Separat test av varje ord i vokabulären

Långsam vid stor vokabulär

- Spektralt distansmått olämpligt

En fonetisk, sannolikhetsbaserad inriktning behövs

GSLT Tal- och talatigenkänning M. Blomberg, 2002-9-9 | 34 |



Sannolikhetsbaserad igenkänning

Bayes' regel för betingade sannolikheter

$$P(\text{Ord} / \text{Akustik}) = \frac{P(\text{Akustik} / \text{Ord}) \times P(\text{Ord})}{P(\text{Akustik})}$$

$P(\text{Ord}/\text{Akustik})$ är *a posteriori sannolikheten* för en ordföljd givet den akustiska informationen.

$P(\text{Akustik}/\text{Ord})$ är *sannolikheten* att ordföljden genererar den akustiska informationen och beräknas i ett träningsmaterial.

$P(\text{Ord})$ ges av språkmodellen och är *a priori sannolikheten* för ordföljden (N-gram).

$P(\text{Akustik})$ kan ses som en konstant eftersom den är oberoende av ordföljden och kan ignoreras

Kombinerar akustisk och språklig kunskap!

GSLT Tal- och talatigenkänning M. Blomberg, 2002-9-9 | 35 |



Dolda Markovmodeller

HMM, Hidden Markov Models



Figure 3.2: Who is this Markov fellow, and why is he hiding?

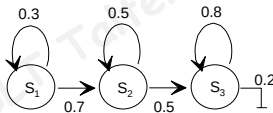
From Bourlard & Morgan
COURSE: SPEECH RECOGNITION - A PRACTICAL APPROACH
1994

GSLT Tal- och talatigenkänning M. Blomberg, 2002-9-9 | 36 |



En Markovprocess

- Modelleras med tillstånd förbundna med bågar med övergångssannolikheter
- Saknar minne
- Förloppet beror enbart på det aktuella tillståndet, ej på historien

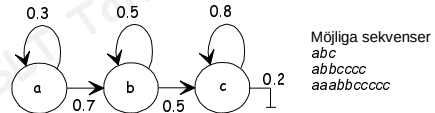


GSLTal- och taligenkänning M Blomberg 2002-9-9 | 37 |



En enkel Markovprocess

- En Markovprocess är en *produktionsprocess*
 - (inte en igenkänningsmetod)
 - processen består av en följd av *tillstånd* (states) som t ex producerar en följd symboler
 - sannolikheten att gå från ett tillstånd till ett annat benämns *övergångssannolikhet*.
 - förloppet beror enbart på det aktuella tillståndet, inte på historien



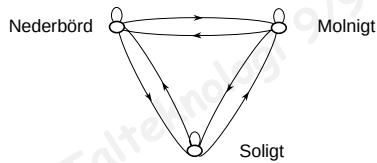
Möjliga sekvenser
abc
abbbccc
aaabbbcccc

GSLTal- och taligenkänning M Blomberg 2002-9-9 | 38 |



Synlig Markovprocess

Exempel: Vädret



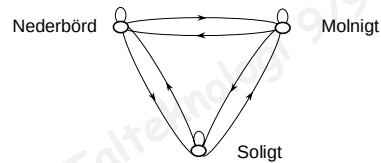
- Processens tillstånd är direkt observerbart

GSLTal- och taligenkänning M Blomberg 2002-9-9 | 39 |



Dold (Hidden) Markovprocess

Exempel: Vädret betraktat genom en gardin



- Processens tillstånd är inte direkt observerbar
 - Man kan göra indirekta, relaterade observationer som ljusstyrka, skarpa skuggor, ljud av regndroppar, lyssna på radio, osv.

GSLTal- och taligenkänning M Blomberg 2002-9-9 | 40 |

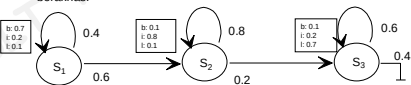


HMM för taligenkänning

- Modell
 - en följd av tillstånd
 - varje modell representerar ett fonem (fonembaserat system) eller ord (ordbaserat)
- Tillstånd
 - delar av fonem (fonembaserat) eller fonem (ordbaserat)
- Övergångssannolikheter mellan tillstånd
 - enkel durationsmodell (exponentialfördelning)
- Observationssannolikheter i tillstånd
 - statistiska modeller av taljudens akustik
 - diskreta (ex. VQ) eller kontinuerliga observationer
- Tillståndsssekvensen kan inte observeras direkt
 - Sannolikheten för att modellen ska ha genererat en följd av observationer kan beräknas.

Diskret
a: 0.3
b: 0.6
c: 0.1
(el. VQ)

Kontinuerlig
 μ : 27.3
 σ : 4.6

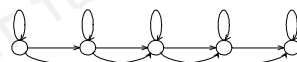


GSLTal- och taligenkänning M Blomberg 2002-9-9 | 41 |



HMM egenskaper

- Snabbare än mönsterigenkänning på grund av sin segmentella struktur
- Möjliggör användande av statistiska metoder
- Fonetiska modeller: snabbare, vokabuläroberoende
 - Lämplig för stora vokabulärer och kontinuerligt tal



GSLTal- och taligenkänning M Blomberg 2002-9-9 | 42 |

Transitionsmatrisen

- Förbindelser mellan tillstånd i en HMM kan representeras i en transitionsmatris som innehåller sannolikheten för övergångarna
- Rad anger från-tillstånd, kolumn anger till-tillstånd

	a	b	c	end
a	0.3	0.7	0.0	0.0
b	0.0	0.5	0.5	0.0
c	0.0	0.0	0.8	0.2
end	0.0	0.0	0.0	1.0

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 [43]

Diskret HMM

- Diskreta utmatade observationer
- Talet kvantiseras med VQ

+ godtycklig statistisk funktion

- kvantiseringsfel,
- VQ och HMM optimeras separat
- kräver mycket träningsmateriel

VQ kodbok kodord-sannolikheter tillstånd

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 [44]

Kontinuerlig HMM

- Kontinuerliga observationer (parametervärden)
- Godtycklig frekvensfunktion med sammanslagning av flera normalfördelningskurvor. (Gaussian mixtures)

+ inget kvantiseringsfel
+ kräver mindre träningsdata om få komponenter

- kräver ofta många normalfördelningskomponenter
- beräkningsintensiv

komponenter blandningsvikt tillstånd

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 [45]

Representation av allofoner

- fonem med varianter (drygt 40 i svenskan)

- Kontextberoende allofoner
 - trifoner, difoner, monofoner beroende på om båda, ena eller inget av de omgivande allofonerna är bestämd
- Varje allofon är ofta uppdelad i flera tillstånd
 - beskriver t. ex. initiala, mediala och finala delar.
- Varje tillstånd bestäms med fördelningar för
 - observationssannolikheter
 - övergångssannolikheter

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 [46]

Sannolikhetsberäkning att en HMM genererat yttrandet

Två algoritmer:

Viterbi:
Sannolikheten är produkten av observationssannolikheter och transitionssannolikheter för **den bästa** tillståndssekvensen

$$P(O|M) = \prod_{t=1}^T P(o_t | s_t) P(s_t | s_{t-1})$$

Maximum likelihood (Baum-Welch):
Sannolikheten beräknas för **alla tillåtna** tillståndssekvenser

$$a_{ij}(t) = P(o_{t+1} | s_i, t) \sum_{k=1}^N a_{jk}(t) P(o_t | s_j)$$

$$P(o_t | s_i) = \sum_{k=1}^K a_{ik}(t)$$

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 [47]

Matchning mellan en HMM-modell och ett yttrande - Viterbisökning

Förenklad algoritm för modellen nedan:

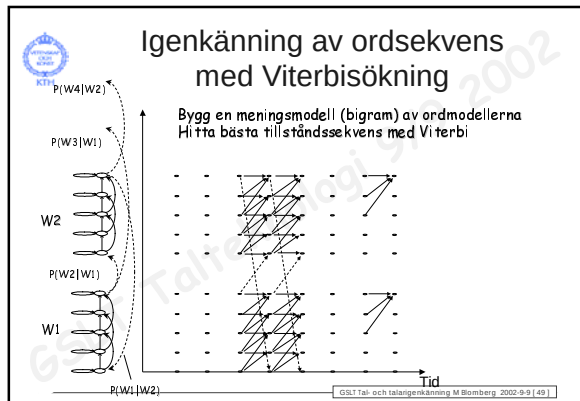
```

for(t=1; t<=T; t++)
  for(s=1; s<=S; s++)
    P(t,s) = P(O_t|S_t) * Max[P(t-1,s)*P_{ss}(s), P(t-1,s-1)*P_{ts}(s|s-1), P(t-1,s-2)*P_{ss}(s|s-2)];
return(P(T,S));

```

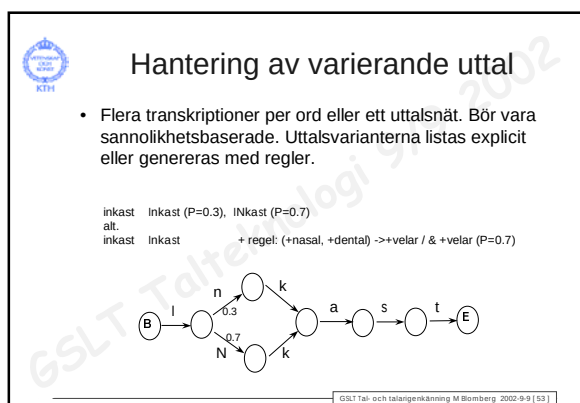
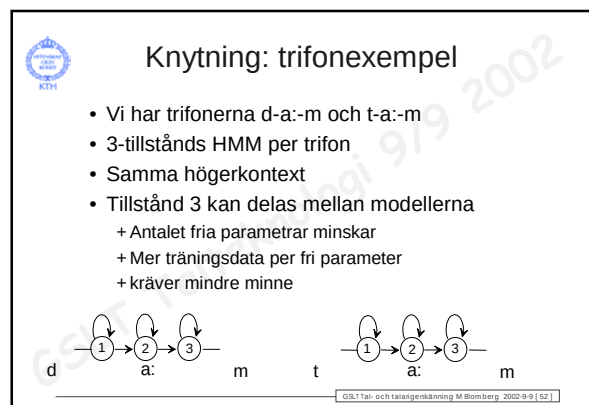
Beg End Tid

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 [48]



- ### Miljontals parametrar
- System med stor vokabulär och ordgränstrifoner använder ~ 60 000 trifoner
 - Akustisk vektor med 39 element och 10 Gausskomponenter (2 parametrar: μ , σ) ger 790 parametrar per tillstånd, varav 10 vikter (mixture weights)
 - $39 \times 10 \times 2 + 10 = 790$
 - 60 000 trifoner med 3 tillstånd ger 142 000 000 parametrar!
 - Knytning (*tying*) för att dela på modellparametrar
 - ger ~10 000 000 parametrar.
- GS<al- och talatigenkänning M Blomberg 2002-9-9 | 50 |

- ### Knutna HMM-system (tied)
- Träningsdata är ojämnt fördelat
 - Exempel: trifoner
 - Det krävs en metod som balanserar modellens komplexitet mot tillgängliga data
 - De olika modellerna får dela på parametrar
 - Automatiska klustringsmetoder används för att optimera sammanknytningen
 - Exempel: beslutsträd
 - slå ihop tillstånd så man får tillräckligt med data per parameter
- GS<al- och talatigenkänning M Blomberg 2002-9-9 | 51 |



- ### Lingvistisk kunskap i taligenkänning
- Finita nät
 - formaliserat språk, t ex flygtrafikledning
 - Unigram, bigram, trigram, n-gram
 - sannolikheter för ordföljder, statistiskt baserade
 - att inkludera djupare lingvistisk kunskap ger ofta bara marginella förbättringar
 - Parsning, språklig analys
 - syntaktiska och semantiska begränsningar
 - Ofta används en tvåstegsprocedure
 - en enkel modell (t ex bigram) används vid den första sökningen för att få ett antal meningsförslag
 - dessa sorteras om med hjälp av mer avancerade språkliga modeller (parsning)
- GS<al- och talatigenkänning M Blomberg 2002-9-9 | 54 |



Vanliga språkmodeller för taligenkänning

- N-gram (ordföljdssannolikheter) ger bra resultat trots sin enkelhet
 - bigram: $P(w_i | w_{i-1})$ *ordpar*
 - trigram: $P(w_i | w_{i-2}, w_{i-1})$ *treordsföljd*
- Ordklassföljder vid otillräckligt träningsmaterial

GSLTal- och taligenkänning M Blomberg 2002-9-9 | 55 |



Kombinera akustisk och språklig kunskap Bayes' regel för betingade sannolikheter

$$P(\text{Ord} | \text{Akustik}) = \frac{P(\text{Akustik} | \text{Ord}) \times P(\text{Ord})}{P(\text{Akustik})}$$

$P(\text{Ord} | \text{Akustik})$ är *a posteriori sannolikheten* för en ordföljd givet den akustiska informationen.

$P(\text{Akustik} | \text{Ord})$ är *sannolikheten* att ordföljden genererar den akustiska informationen och beräknas i ett träningsmaterial.

$P(\text{Ord})$ ges av språkmodellen och är *a priori sannolikheten* för ordföljden (N-gram).

$P(\text{Akustik})$ kan ses som en *konstant* eftersom den är oberoende av ordföljden och kan ignoreras

GSLTal- och taligenkänning M Blomberg 2002-9-9 | 56 |



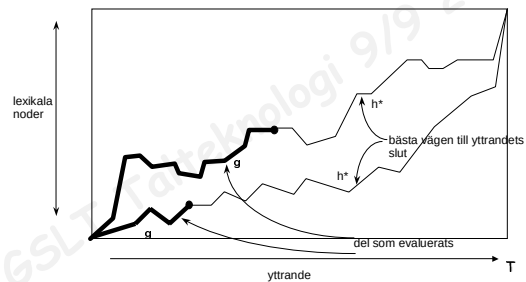
N-bästa sökning med A*

- Utför en "Best First" sökning med flera hypoteser
- Den första som täcker yttrandet är bäst, andra näst bäst osv.
- Problem: Monoton sjunkande sannolikhet vid ökande hypoteslängd => korta hypoteser gynnas -> snarlik längd -> långsamt
- Lösning: Lägg till varje hypotes' maximala sannolikhet (h^*) för den ej sökta delen av yttrandet -> alla hypoteser får samma längd.
- Metrik: $P(\text{hyp}, t) = g[0, t] + h^*[t+1, T]$ (mätt + uppskattad)
- h^* för alla tider och tillstånd erhålls med normal Viterbi-sökning i ett inledande pass (motsatt riktning).

GSLTal- och taligenkänning M Blomberg 2002-9-9 | 57 |



A*-heuristik i sökmatrisen



GSLTal- och taligenkänning M Blomberg 2002-9-9 | 58 |



N-bästa exempel de 10 akustiskt bästa meningarna

Nr	Akustisk poäng	Igenkänd mening
1	-716,914	hur åka till svartlöga
2	-717,444	du åka till svartlöga
3	-718,289	ju åka till svartlöga
4	-719,614	hur åker man till svartlöga
5	-719,730	hur åka av hamn till svartlöga
6	-720,260	du åka av hamn till svartlöga
7	-720,365	ut åka till svartlöga
8	-720,554	hur båtarna till svartlöga
9	-720,630	hur åker hamn till svartlöga
10	-720,699	nu åka till svartlöga

GSLTal- och taligenkänning M Blomberg 2002-9-9 | 59 |



Prestanda för taligenkänning

- Viktiga faktorer
 - vokabulärstorlek
 - språkmodell (perplexitet)
 - talarberoende / talarberoende
 - akustisk miljö
 - kanalens egenskaper
 - talarkarakteristik ("får och getter")
 - ingående ords akustiska likhet

GSLTal- och taligenkänning M Blomberg 2002-9-9 | 60 |



Utvärdering av taligenkänning

$$\text{Ordnoggrannhet} = \frac{100 * (\text{Antal_ord} - \text{Förväxlingar} - \text{Instopp} - \text{Bortfall})}{\text{Antal_ord}}$$

Om tidsgränser för orden i den rätta meningen saknas kan en garanterat korrekt bedömning inte göras. *Dynamisk programmering* kan i så fall användas för att synkronisera orden mellan den igenkända och korrekta meningen. Detta ger en optimal matching som inte behöver motsvara det verkliga igenkänningsresultatet.

Stim	Resp	Jag	vill	åka	idag	till	Vaxholm
Jag	Korrekt						
skulle	Bortfall						
villja	Förväxling						
åka	Korrekt						
till	Instopp						
Vaxholm	Korrekt						

6 ord, 1 förväxling, 1 instopp, 1 bortfall: Ordnoggrannhet = $100 * (6 - 3) / 6 = 50\%$

Tar ej hänsyn till fonetisk likhet mellan ord

Ex: i kväll - ikväll, jag - ja, Vaxholm - Vaxholms, bil - restaurang är alla lika mycket fel

GSLT Tal- och taligenkänning M. Blomberg, 2002-9-9 | 61 |



Tillämpningens svårighet - språkets perplexitet

$$H = - \sum_{\forall W} P(W)^2 \log P(W)$$

$$B = 2^H$$

B perplexiteten för givna tillämpningen

H entropin för den givna tillämpningen

$P(W)$ sannolikheten för ordsekvensen W i det givna språket

Perplexiteten ger ett mått på det sannolikhetsviktade genomsnittliga antalet ord som kan följa efter ett givet ord.

Exempel: För siffersekvenser är $B = 10$, om alla siffror är lika sannolika i alla positioner. Vid ojämn sannolikhetsfördelning är $B < 10$.

Måttet tar ingen hänsyn till den fonetiska skillnaden mellan olika ord.

GSLT Tal- och taligenkänning M. Blomberg, 2002-9-9 | 62 |



ARPA-projektet i USA

- Advanced Projects Research Agency
 - började 1984
 - deltagare
 - CMU, SRI, BBN, MIT, Lincoln Labs, Dragon Systems
 - "competitive evaluations" varje år
 - domäner
 - ATIS, Air Traffic Information System, ~1000 ord
 - Flygbokning
 - Wall Street Journal, 5000 - 60000 ord
 - Tidningstext, uppläst
 - SWITCHBOARD
 - Samtal över telefon med okänd person om givet ämne
 - CALL HOME
 - Samtal över telefon med närmaste familjekretsen
 - Broadcast News
 - radionyheter inklusive telefonintervjuer, jinglar med mera

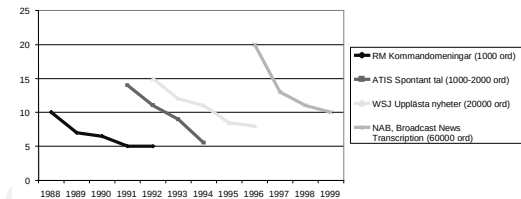
svårare

GSLT Tal- och taligenkänning M. Blomberg, 2002-9-9 | 63 |



Prestandautveckling DARPA-utvärdering 1988-1999

Ordfel (%)

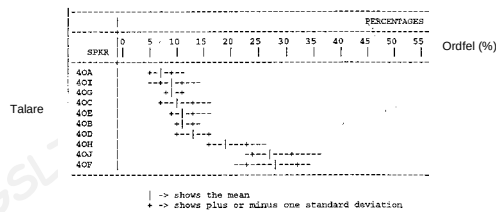


GSLT Tal- och taligenkänning M. Blomberg, 2002-9-9 | 64 |



Får och getter ("bra" och "dåliga" talare)

ARPA-utvärdering, Wall Street Journal
nov 1993 Hub1

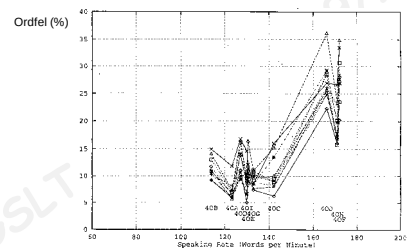


GSLT Tal- och taligenkänning M. Blomberg, 2002-9-9 | 65 |



Talshastighetens inverkan

ARPA-utvärdering, Wall Street Journal
nov 1993 Hub1



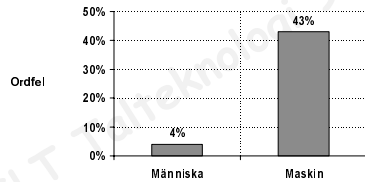
GSLT Tal- och taligenkänning M. Blomberg, 2002-9-9 | 66 |



Jämförelse människa -maskin

Resultat för mänsklig transkription av hela samtal jämfört med automatisk igenkänning av extraherade meningar från samma material.

Talmaterial: Telefonsamtal om kreditkort i databasen "Switchboard".



R. P. Lippman: "Speech perception by humans and machines", Proc. of the workshop on the auditory basis of speech perception, Keele University, UK, 15-19 July, 1996, pp. 309-316.

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 (67)



Begränsningar i dagens system

- Stora mängder tal behövs för inläring
 - man måste veta vad som sagts
- Talaroberoende system
 - tar ingen hänsyn till att talaren är "konstant"
 - alla ramar i ett och samma yttrande kan ha producerats av olika talare
- Prosodisk information utnyttjas inte
 - inte segmentell
 - hur ska den mätas?
- Begränsad intelligens och omvärldsförståelse
 - men fungerar inom begränsade domäner

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 (68)



IEEE Workshop on ASR, Snowbird, Dec. 1995

Människan och maskinen

Resultat av omröstning: "Om hur många år får maskinen samma resultat som människan för akustisk/fonetisk igenkänning?"

60 talforskare svarade:

Antal år	Antal röster
5 - 10	10
10 - 20	30
20 - 40	10
> 40	10

GSLT Tal- och talatigenkänning M Blomberg 2002-9-9 (69)