Ein- und Ausgabe von Sprache

0525250 Christoph Redl[1]

Wintersemester 2008/2009
Version 1.1, 10.12.2008

[1]E-mail: e0525250@mail.student.tuwien.ac.at

Inhaltsverzeichnis

Vorwort

Dieses inoffizielle Skriptum zur Vorlesung „3.0/2.0 VO Ein- und Ausgabe von Sprache“ von Dr. Markus Kommenda im Wintersemester 2008/09 wurde von Christoph Redl erstellt. Es wurde nach bestem Wissen und Gewissen verfasst, vor allem um den Stoff für mich selbst zu wiederholen und zu vertiefen. Dennoch erhebt es keinen Anspruch auf Richtigkeit und Vollständigkeit. Die Inhalte wurden von der LVA-Leitung nicht überprüft.

Die Quellen, auf denen diese Mitschrift aufbaut, sind am Ende angeführt. Im Wesentlichen stützt sich der Inhalt aber auf die Vorlesung und die zugehörigen Folien.

Kapitel 1

Sprachwissenschaftliche Grundlagen

1.1 Einführung

Die Aufgabestellungen in der digitalen Sprachverarbeitung sind vielseitig. Grundsätzlich kann unterschieden werden, zwischen wem Kommunikation stattfindet. Kommunizieren zwei oder mehrere Menschen miteinander, so handelt es sich um reine Sprachsignalübertragung. Ist jedoch einer der Kommunikationspartner eine Maschine, so kommt Sprachsynthese bzw. Sprachwiedergabe, oder Spracherkennung (manchmal auch Sprechererkennung, z.B. bei Sicherheitssystemen) ins Spiel. Kombiniert man beides mit einer Verarbeitungslogik dazwischen, so spricht man von Sprachdialogsystemen.

Bei der Ausgabe von Sprache kommt es auf die Anwendung an, ob die Wiedergabe von vorgesprochenen Textstücken, oder Sprachsynthese, also die künstliche Erzeugung beliebiger Texte, verwendet wird. Beispiele für reine Sprachwiedergabe sind die elektronische Zeitansage, bei der die Uhrzeit aus wenigen vorgesprochenen Textbausteinen (z.B. die Ziffern) zusammengesetzt wird, oder auch die Stationsansagen in U-Bahn und Straßenbahn. Sollen hingegen beliebige Texte elektronisch wiedergegeben werden können, etwa bei Screenreadern für Blinde, so muss man auf Sprachsynthese zurückgreifen. Auch dabei werden vorgesprochene Texte verwendet, allerdings nicht auf Wortgruppen- oder Wortebene, sondern auf Lautebene.

Grundsätzlich lässt sich jedes Sprachsystem mit folgendem Schema beschreiben:

1. Ein akustisches Signal wird von einem Mikrofon erfasst
2. Es findet eine Signalwandlung statt (in ein elektronisches Signal)

3. Das Signal wird zum System übertragen
4. Spracherkennung
5. Nicht-sprachliche Verarbeitung
6. Sprachausgabe
7. Übertragung des elektronischen Signals zum Ausgabemedium
8. Signalwandlung (in ein akustisches Signal)
9. Es entsteht wieder ein hörbares Sprachsignal

Je nach System können natürlich auch Teilschritte fehlen oder neben der Verarbeitung auch eine Speicherkomponente eingefügt werden.

Anwendungsbeispiele:

- Ansagen in öffentlichen Verkehrsmittln oder auf Bahnhöfen
- Screenreader
- Sprachwahl beim Handy
- Elektronische Zeitansage
- Diktiersoftware

1.2 Grundbegriffe aus der Sprachwissenschaft

1.2.1 Anatomie

Die Sprachproduktion erfolgt beim Menschen durch verschiedenste Organe. Dabei lassen sich 3 Hauptgruppen herauszeichnen.

- Die **Respiration** (Atmung) mit Hilfe der Lunge und der Brust- und Bauchmuskulatur.
- Die **Phonation** (Stimmgebung) durch den Kehlkopf mit den Stimmbändern
- Die **Artikulation** (Lautbildung) im Mund-, Nasen- und Rachenraum, durch die Lippen, die Zunge und die Zähne, den Zahndamm (alveolar), den harten und weichen Gaumen (palatum und velum), das Gaumensegel und das Zäpfchen. Bewegliche Teile werden als aktiv, die anderen als passiv bezeichnet.

1.2.2 Laute und deren Bildung

Die vom Menschen erzeugten Laute lassen sich bekanntlich in Vokale und Konsonanten einteilen. Unterscheidungsmerkmal ist, dass bei Vokalen der Mundraum eher weit geöffnet ist und die Luft relativ ungehindert hinausströmen kann. Außerdem sind Vokale in der Regel immer stimmhaft (außer beim Flüstern), während es auch stimmlose Konsonanten gibt.

Als Vokale gelten nicht nur die üblichen fünf: a, e, i, o und u, sondern auch noch die Umlaute und Zwielaute. Außerdem gibt es von den zuvor genannten auch mehrere Abstufungen. Um das zu verstehen ist es notwendig einzusehen, dass Laute und Buchstaben nicht das gleiche sind. Wie man sich leicht anhand konkreter Wörter überlegen kann, wird etwa ein e nicht immer gleich ausgesprochen: manchmal näher am ä, manchmal näher am i. Je nach Sprache lassen sich verschiedene Anzahlen von Vokalen unterscheiden.

Bei Konsonenten verhält es sich ähnlich. Auch hier sind Buchstaben und Laute nicht äquivalent. Die IPA (International Phonetic Association) hat deshalb klar definiert, welche Laute wie gebildet werden (durch Röntgenbilder der beteiligten Organe untermauert), und wie diese Laute in der Lautschrift geschrieben werden (sowie die Schreibweisen mit dem relativ beschränkten ASCII-Zeichensatz). Besonders wichtig ist diese Definition beim Erlernen von Fremdsprachen.

Die Vokale lassen sich gut in das Trapez in Abbildung 1.1, dem sogenannten **Vokalviereck**, einzeichnen. Die Abszisse gibt dabei die horizontale Zungenposition und die Ordinate die vertikale an. Man erkennt somit, dass beispielsweise bei einem i: die Zunge vorne weit oben ist, während sie bei einem u: eher hinten nach oben geht.

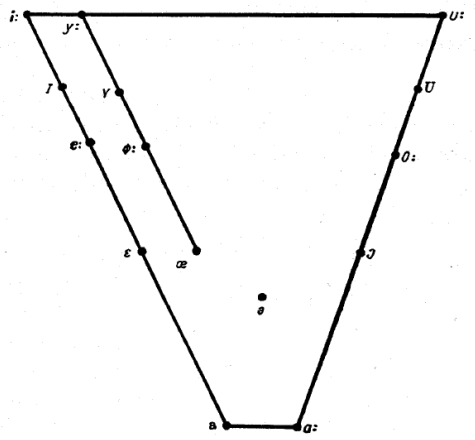

Abbildung 1.1: Das Vokalviereck

Bei Konsonanten lassen sich folgende Kriterien unterscheiden:

- **Stimmhaftigkeit**: Je nachdem ob die Stimmbänder mitschwingen oder nicht
- **Artikulationsart**: Je nachdem ob ein Laut durch Versperrung des Luftweges und dann plötzliche Öffnung (Verschlusslaute, Plosive), durch Reibung (Frikative, Reibelaute) oder über die Nase (Nasale) gebildet werden.
- **Artikulationsort**: Je nachdem wo die Verengung oder Versperrung des Luftweges auftritt: bei den Lippen (labial), bei den Zähnen (dental), am Zahndamm (alveolar), am harten Gaumen (palatal), am weichen Gaumen (velar), beim Zäpfchen (uvular) oder bei der Stimmritze (glottal)

Beispiele:

- Das [d] ist ein stimmhafter, alveolarer Verschlusslaut (Plosiv)
- Das [s] ist ein stimmloser, alveolarer Reibelaut
- Das [m] ist ein bilabialer Nasal

1.2.3 Phonetik und Phonologie

Es werden nicht nur Buchstaben oft unterschiedlich ausgesprochen, es werden sogar die gleichen logischen Laute, etwa ein [d], nicht immer gleich gesprochen. Das Gesprochene im Sinne der Akustik unterscheidet sich stark von Sprecher zu Sprecher, und auch eine einzelne Person klingt nicht immer exakt gleich. Dennoch steht jede dieser unterschiedlichen Ausprägungen für ein und denselben Laut.

Die konkrete Ausprägung wird als **Phon** bezeichnet, während die logischen Klassen, in die man die Phone einteilen kann, **Phoneme** genannt werden.

Anmerkung: Es hängt durchaus von der Sprache ab, ob mehrere Lautausprägungen Phone oder Phoneme sind. In manchen Sprachen wird etwa zwischen dem uvularen und dem Zungenspitzen-R unterschieden, während im Deutschen beides erlaubt ist und dieselbe Bedeutung hat. Unterscheiden sich zwei oder mehrere Wörter nur durch einen einzigen Laut, so spricht man von einer **distinktiven** Funktion dieses einen Lautes.

In diesem Zusammenhang ist zu erwähnen, dass das Wort „Sprache“ im Deutschen eine Doppelbedeutung hat, nämlich einerseits die Sprache

als System und andererseits die Sprache auf physikalischer Ebene, in Form von Schallwellen. In vielen Sprachen gibt es dafür unterschiedliche Wörter, beispielsweise in Englisch: *language* für die Sprache als System und *speech* für die gesprochene Sprache.

1.2.4 Graph und Graphem

Analog zu Phon und Phonem, bezeichnet **Graph** in der geschriebenen Sprache eine konkrete Ausprägung eines Buchstabens (Schriftart, Druck-/Handschrift, fett, kursiv, etc.), während ein **Graphem** die logische Einheit des Buchstabens ist, z.B. das a.

1.2.5 Morph und Morphem

Ein **Morphem** ist die kleinste eigenständig sinntragende Einheit in einer Sprache. Dazu gehören Wortstämme als freie Morpheme und Prä- oder Suffixe als gebundene Morpheme, etwa: -lich, -kein, ver-, usw.. Ein **Morph** ist wiederum eine konkrete Ausprägung eines Morphems. Damit werden etwa Singular und Plural eines Wortes, beides das gleiche Morphem, unterschieden.

Anmerkung: Morpheme fallen nicht immer mit Silben zusammen. Eine Silbe ist eine Gruppe von Lauten, in der mindestens ein Vokal vorkommt. Vor und nach dem Vokal (dem Inlaut) kann (aber muss nicht) noch eine Gruppe von Konsonanten oder weiteren Vokalen stehen (Anlaut und Auslaut).

1.2.6 Ambiguität

Die Sprache ist nicht eindeutig. Mehrdeutigkeit kann auf **lexikalischer Ebene** (mehrere Bedeutungen eines Wortes), auf **Ebene der Wortart** (Unklarheit darüber, als welcher Worttyp ein Wort in einem konkreten Satz fungiert) oder auf **struktureller syntaktischer Ebene** (Unklarheit, wie ein Satz - z.B. aufgrund mangelnder Interpunktion - zu lesen ist) auftreten.

1.2.7 Prosodie

Die Prosodie bezeichnet sogenannte suprasegmentale Betonungen, also Fragen der Aussprache, die sich über mehrere Laute, Silben oder Wörter erstrecken. Damit können etwa Fragen von Aussagen unterschieden, Sarkasmus erkannt oder Wortgrenzen angedeutet werden (letztere werden in der Regel nicht anhand von Pausen erkannt!).

Die Prosodie gewinnt auch in der Sprachsignalverarbeitung zunehmend an Bedeutung und ist derzeit ein wesentlicher Grund dafür, dass syntheti-

sierte Sprache oft noch immer künstlich klingt.

Prosodische Parameter sind etwa die Lautstärke, die Sprachgrundfrequenz, die Lautdauern und die Pausen. Beeinflusst werden all diese Parameter von den Intonationsmotiven des Sprechers, dem Sprechrhytmus und Akzenten.

Kapitel 2

Sprachsignalverarbeitung

2.1 Grundlagen der Akustik

Ein Sprachsignal, oder allgemeiner: ein akustisches Signal, kann grundsätzlich im Zeit- oder Frequenzbereich angegeben werden, wobei mit Hilfe geeigneter Transformationen, etwa der DCT (Diskrete Cosins-Transformation), zwischen beiden transformiert werden kann.

2.1.1 Zeitbereich

Gibt man ein Signal im Zeitbereich an (siehe Abbildung 2.1), so wird in Abhängigkeit von der Zeit t auf der Ordinate der Luftdruck angegeben. Es ist erkennbar, dass Vokale grundsätzlich aufgrund der Schwingung der Stimmbänder eine gewisse Periodizität aufweisen. Stimmlose Konsonanten gleichen dagegen eher einem Rauschen. Selbstverständlich sind aber auch Vokale bzw. stimmhafter Konsonanten keine reine Sinusschwingung, sondern eine Überlagerung einer Grundfrequenz mit ihren Obertönen (ganzzahlige Vielfache der Grundfrequenz) sowie gewisser Rauschanteile. Zum Vergleich siehe Abbildung 2.2.

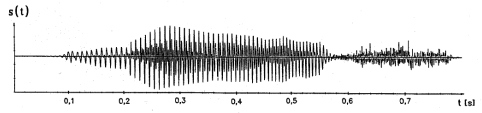

Abbildung 2.1: Das Signal im Zeitbereich

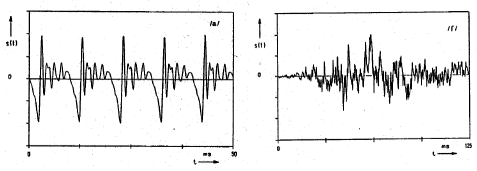

Abbildung 2.2: links: Vokal oder stimmhafter Konsonant, rechts: stimmloser Konsonant

2.1.2 Frequenzbereich

Im Unterschied dazu wird im Frequenzbereich (siehe Abbildung 2.3) für jede Frequenz angegeben, wie stark diese im akustischen Signal enthalten ist. Durch Überlagerung vieler Sinusschwingungen mit unterschiedlichen Frequenzen kann jedes akustische Signal beschrieben werden.

Es ist anzumerken, dass im Frequenzbereich - anders als im Zeitbereich - immer ein Zeitfenster und kein Zeitpunkt betrachtet wird. Das Frequenzspektrum bezieht sich also immer auf einen Zeitraum.

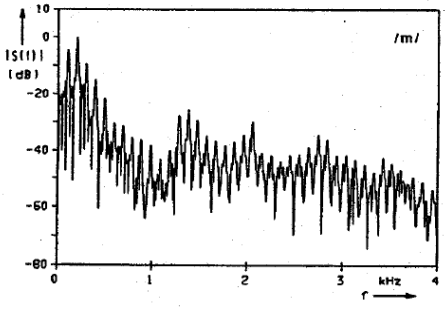

Abbildung 2.3: Das Signal im Frequenzbereich

2.1.3 Sonagramme

Es gibt dennoch eine Möglichkeit das Signal gleichzeitig im Zeit- und Frequenzbereich zu betrachten. Dazu wird ein sogenanntes **Sonagramm** erstellt, in dem entlang der Abszisse die Zeit, und entlang der Ordinate die Frequenz aufgetragen wird. Eine Schwärzung in einem bestimmten Punkt gibt dabei nun an, wie stark zu einem konkreten Zeitpunkt eine bestimmte Frequenz im Signal enthalten war. Die Stärke wird dabei als 3. Dimension in Form des Schwärzungsgrades codiert.

Je nachdem ob man das Diagramm in zeitlicher oder Frquenz-Dimension

feiner darstellt, spricht man von einem Breitband- oder Schmalbandsonagramm.

Im Breitbandsonagramm verschwimmen die Frequenzen sehr stark, dafür kann man aber die Formanten deutlich erkennen. Das sind „größere schwarze Blöcke“, die die Energiekonzentrationen des fertigen Sprachsignals angeben. Anders gesagt: zu welchen Zeitpunkten auf welchen Frequenzbereichen etwas hörbar war. Außerdem sind bei stimmhaften Lauten die Schläge der Glottis in Form von harmonischen Senkrechten mit vielen lokalen Maxima über den gesamten Frequenzbereich erkennbar. Bei stimmlosen Lauten fehlen genau diese lokalen Maxima und die Energie ist gleichmäßiger verteilt, siehe dazu den ersten Abschnitt in Abbildung 2.6. Diese Striche kommen dadurch zustande, dass sich die Grundfrequenz von etwa 100 Hz (bei einem männlichen Sprecher) mit ihren zahlreichen Obertönen überlagert. Wegen der geringen Auflösung in Freuquenzichtung sind die Schläge nur als zusammenhängende schwarze Striche sichtbar.

Dagegen verschwimmen im Schmalbandsonagramm die Zeiten sehr stark, wegen der feinen Auflösung der Frequenzen treten dafür die Grund- und Obertöne deutlich hervor.

Zum Vergleich siehe Abbildung 2.5.

Die Formanten entstehen durch die Resonanzeigenschaften im Artikulationstrakt (Mund- und Rachenraum). Durch die dort herrschenden Verengungen werden die Resonanzfrequenzen beeinflusst, und so gewisse Frequenzbereiche relativ zu den anderen verstärkt und andere geschwächt. Bei einem komplett geöffneten Mund mit der Zunge (Anmerkung: dies entspricht dem Laut [a]) ganz unten ergibt sich eine Resonanzfrequenz von 500 Hz (die aus der Schallgeschwindigkeit und der Länge des Mund- und Rachenraumes ermittelt werden kann), sowie weiteren Resonanzfrequenzen als ungerade ganzzahlige Vielfache davon. Daraus ergibt sich auch, dass gerade diese Frequenzbereiche bei einem gesprochenen [a] besonders stark ausgeprägt sind, was als Formanten sichtbar wird.

Erfolgt nun eine Verengung im Schallschnellenbereich, so vermindert sich die Resonanzfrequenz, erfolgt sie im Beeich des maximalen Luftdrucks, so erhöht sie sich. Dadurch werden andere Frequenzbereiche des gesamten (von der Glottis) erzeugten Spektrums „durchgelassen“, oder anders gesagt: **herausgefiltert**. Dies führt zum **Quelle-Filter-Modell**.

Das Vokalviereck kann mit dieser Sichtweise nun - mit der gleichen geometrischen Form - als **akustisches Viereck** (siehe Abbildung 2.7) bezeichnet werden. Die zwei Dimensionen entsprechen den ersten beiden Resonanz-

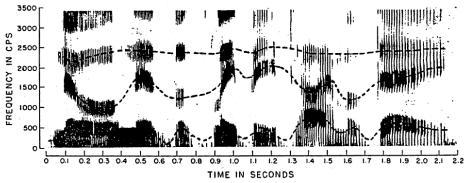

Abbildung 2.4: Das Signal im Zeit- und Frequenzbereich

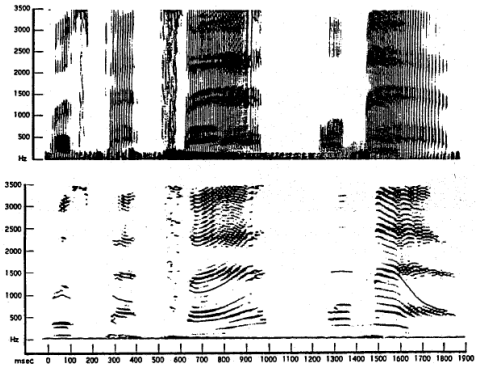

Abbildung 2.5: Breit- (oben) und Schmalbandsonagramm (unten)

frequenzen, oder äquivalent: der Lage der ersten beiden Formanten. Genau diese Formantenlage kann zur Lauterkennung verwertet werden.

2.2 Grundlagen der Spracherkennung

Schlüssel zur Erkennung gesprochener Laute sind die Lagen der sogenannten **Formanten** im Sonagramm. Formanten bezeichnen dabei größere geschwärzte Blöcke, also Frequenzbereiche, die besonders stark vertreten sind, wenn ein Laut gesprochen wird. Besonders der erste und zweite (nach aufsteigenden Frequenzen) Formant ist sehr aussagekräftig.

Das Vokalviereck (siehe Abbildung 1.1) ist auch bei der Lauterkennung hilfreich: zeichnet man die Lage des ersten Formanten von rechts nach links, und die des zweiten von oben nach unten (Achsen in Richtung der steigenden Frequenzen) auf, so sieht das Vokalviereck ganz ähnlich aus, und wird in diesem Zusammenhang als **akustisches Viereck** bezeichnet.

Problematisch bleibt dabei lediglich, dass die Lage der einzelnen Formanten weder bei einem Sprecher, und schon gar nicht bei verschiedenen Sprechern, gleichbleibend ist. Besonders schlecht ist jedoch, dass sich die verschiedenen Phone ein- und desselben Phonems gar nicht so eindeutig

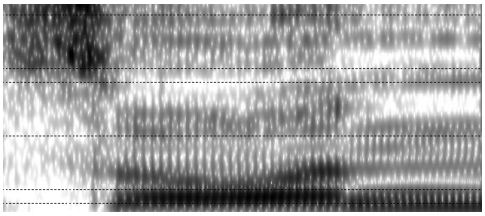

Abbildung 2.6: Stimmhafte und stimmlose Laute im Sonagramm

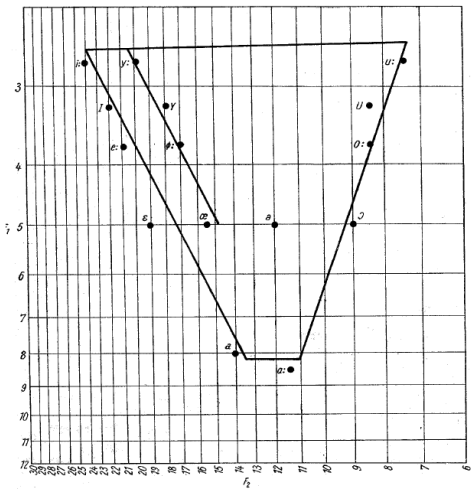

Abbildung 2.7: Akustisches Viereck

klassifizieren lassen, da die Klassen ineinander übergehen. Hier sind gute Näherungen gefragt, die etwa mit neuralen Netzen erzielt werden können.

2.3 Grundlagen der Sprachsynthese

Bei der Sprachsynthese wird ein Modell verwendet, das auch der natürlichen Sprachproduktion beim Menschen entspricht: das **Quelle-Filter-Modell** (siehe Abbildung 2.8).

Es wird von einem Signalgenerator für stimmhafte Laute ausgegangen (beim Menschen durch die Schwingung der Stimmbänder) sowie einem zweiten Generator für Rauschen (entspricht den Luftturbulenzen bei Frikativen). Diese beiden Signale werden überlagert, wobei je nach zu produzierendem Laut das eine oder andere Teilsignal überwiegen kann. Das Gesamtsignal kann zusätzlich auch noch skaliert werden.

Im Anschluss daran wird ein Filter geschaltet, der ein Modell des Vokal-

traktes, also des Mund-, Nasen- und Rachenraumes darstellt. Von diesem Filter wird das davor erzeugte Sprachsignal frequenzabhängig stärker oder schwächer durchgeschleust oder ausgefiltert. Daraus entsteht das Ausgangssignal.

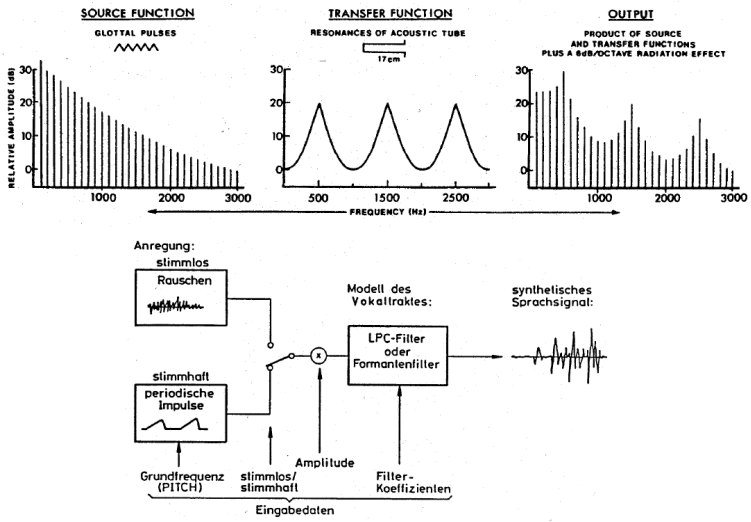

Abbildung 2.8: Quelle-Filter-Modell

Anmerkung: Das Modell ist nicht nur bei der Sprachsynthese, sondern auch bei der Spracherkennung von Bedeutung.

Kapitel 3

Sprachausgabe

3.1 Historisches

Die ersten Versuche zur künstlichen Sprachproduktion wurden 1791 von Wolfgang von Kempelen angestellt. Zu diesem Zweck konstruierte er eine Maschine bestehend aus verschiedenen Blasebalgen, Pfeifen (für Frikative) und Hohlräumen, die den menschlichen Rachenraum nachahmen sollten. Die Maschine konnte, wenn gleich auch mit eher geringer Qualität, einigermaßen verständliche Wörter produzieren. Es war somit gezeigt, dass künstliche Sprachproduktion prinzipiell möglich ist.

Bedeutende Fortschritte konnten erst durch elektronische Modelle erzielt werden. Bei der Weltausstellung 1939 zeigte Homer Dudley seinen *Voice Coder* (kurz: Vocoder oder Voder), der bereits auf das Quelle-Filter-Modell (siehe Abschnitt 2.3) aufbaute. Mit Pedalen konnte stimmhaft/stimmlos unterschieden und im stimmhaften Fall die Grundfrequenz gesteuert werden.

Weitere Fortschritte gab es dann erst wieder mit Aufkommen von leistungsfähiger Elektronik in Form von dedizierten Chips.

3.2 Anwendungen

Grundsätzliches Ziel jeder Sprachausgabe ist die Verbesserung der Benutzerschnittstelle. Die Motive dafür können unterschiedlich sein:

- Annäherung der Interaktion mit einem Rechner an die **menschliche Form der Kommunikation**
- **Frei machen der Hände**, um diese für andere Aufgaben nutzen zu können
- Unterstützung für **Sehbehinderte**

- Automatische Dialogführung über das **Telefon**
- schnelleres **Erwecken der Aufmerksamkeit**, z.B. in Gefahrensituationen
- **Miniaturisierung** von Mobilgeräten, z.B. Handys

Konkrete Beispiele sind Flugzeugcockpits, Lagerverwaltung, Autos, Automatische Bestellsysteme von Versandhäusern, Blindenlesegeräte, Ergänzung der Bildschirmausgabe über einen zusätzlichen Kommunikationskanal, Rechenzentren, etc..

3.3 Typen

Es lässt sich grundsätzlich unterscheiden zwischen **reiner Sprachwiedergabe** und **Sprachsynthese**.

Sprachwiedergabe meint das reine Abspielen von vorgesprochenen Aufzeichnungen, wobei das dynamische Zusammensetzen von größeren Sprachbausteinen möglich ist (etwa die Stationsansage in der U-Bahn oder die Zeitansage). Sprachsynthese erlaubt dagegen die Ausgabe beliebiger Texte aus einem unbeschränkten Vokabular. Es liegt in der Natur der Sache, dass bei Sprachwiedergabe der Sprecher meist erkennbar ist, bei Sprachsynthese hingegen nicht.

Sprachwiedergabe ist wesentlich einfacher zu realisieren, da hier die Prosodie in den von Menschen vorgesprochenen Textblöcken enthalten ist. Synthetisierte Sprache erfordert dagegen ein künstliches Erzeugen der Prosodie, da sie aufgrund der (viel feiner aufgelösten) Textbausteinen, etwa einzelnen Lauten, nicht bereits bei der Aufnahme berücksichtigt werden kann. Dazu sind die Laute viel zu kontextabhängig.

Eine verfehlte Prosodie macht synthetisierte Sprache deutlich als solche erkennbar. Es stimmen beispielsweise Pausen, Betonungen, Lautdauern und der Rythmus nicht. Untersuchungen haben aber gezeigt, dass eine als künstlich erkennbare Sprache nicht automatisch mangelnde Akzeptanz von Seiten des Users bedeutet. Gute Verständlichkeit spielt unter Umständen eine wesentlichere Rolle als Natürlichkeit.

Als Vorteil kann bei der Sprachsynthese, neben dem unbeschränkten Vokabular, noch genannt werden, dass der Speicherplatzbedarf wesentlich geringer ist, da die Texte bis vor der Synthese auch wirklich als Text und nicht als Audiosignal gespeichert werden können. Dennoch ist die Sprachsynthese derzeit noch eher Gegenstand der Forschung, auch wenn bereits einige ganz gute Systeme existieren.

3.4 Sprachverarbeitung

Wird ein Sprachsignal übertragen oder - in dieser LVA vor allem - gespeichert, so lässt sich grundsätzlich zwischen zwei Varianten unterscheiden. Zum einen kann die Zeit-Luftdruck-Funktion direkt im **Zeitbereich** abgetastet, quantisiert und übertragen werden. Man spricht auch von der Signalformcodierung. Alternativ dazu kann das Signal diversen Transformationen unterworfen und **Parameter** des Signals herausextrahiert werden, die dann übertragen werden.

Die reine Sprachausgabe lässt sich durch direkte Anwendung der in den Unterabschnitten 3.4.1, 3.4.2 und 3.4.3 beschriebenen Verfahren umsetzen. Für die Sprachsynthese sind weiterführende Überlegungen notwendig, die in Unterabschnitt 3.4.4 dargestellt werden.

3.4.1 Signalformcodierung im Zeitbereich

Grob umrissen wird zu diskreten Zeitpunkten der Luftdruck als analoger Wert gemessen, auf einen benachbarten darstellbaren digitalen Wert gerundet (Quantisierung), und dieser Wert schließlich gespeichert.

Die Abtastfrequenz muss dabei nach dem Abtasttheorem von Shannon mindestens doppelt so hoch sein wie die höchste vorkommende Frequenz im Signal, damit das Ursprungssignal wieder exakt rekonstruiert werden kann. Wird das Theorem verletzt, so „verschwinden“ Frequenzen nicht einfach, sondern - was noch viel schlimmer ist - es treten falsche Frequenzen auf (das wird als „Aliasing“ bezeichnet). Aus diesem Grund wird vor der Abtastung auch immer ein Tiefpassfilter vorgeschaltet, der zu hohe Frequenzen von vornherein entfernt und das Signal auf das vorgesehene Frequenzspektrum beschränkt.

Die Quantisierung erfolgt streng genommen nach der Abtastung, in der Praxis wird aber oft beides in einem Arbeitsschritt erledigt. Die Qualität ist dabei natürlich umso größer, je mehr (und daher feinere) Stufen es bei der Quantisierung gibt. Der Quantisierungsfehler wird als *Quantisierungsrauschen* hörbar und beträgt bei jedem Sample maximal die halbe Stufenhöhe (in beide Richtungen).

In diesem Zusammenhang ist das **Signal-to-noise-ratio** (SNR) zu erwähnen. Eine Faustregel besagt, dass $SNR = SNR_0 + 6 * Bitanzahl[dB]$, wobei SNR_0 von der Art des Signals abhängt (bei Sprache ungefähr -5 dB).

Konkrete Modulationsverfahren bauen genau auf diesem Prinzip auf. In der einfachsten Ausführung werden die einzelnen quantisierten Samples

einfach übertragen: **linear Pulse-Code-Modulation** (linPCM). Verbesserungen können durch eine logarithmische Skala erzielt werden, bei dem leisere Signale mit höherer Genauigkeit übertragen werden, um einem zu stark singenden SNR entgegenzuwirken (**logPCM**).

Weitere Verbesserungsstrategien bestehen in der Übertragung der (erwartungsgemäß geringeren) Differenzen zwischen 2 Samples statt der Absolutwerte: **Differential PCM** (DPCM). Die Differenzen werden aber ebenfalls quantisiert. Treibt man diese Überlegung an die Spitze, so erhält man die **Deltamodulation** (DM), bei der überhaupt nur noch übertragen wird, ob ein Sample im Vergleich zum vorherigen größer oder kleiner geworden ist. Die Abstufung ist nur noch hardcoded bzw. wird in der Programmlogik mitgeführt.

Adaptive PCM (APCM) baut auf DPCM auf und verwendet abhängig von dem für das nächste Zeitintervall geschätzten Signalunterschied unterschiedliche Bitanzahlen für die Differenzcodierung. So wird etwa bei einem geringen Unterschied eine geringere Anzahl von Bits verwendet, während bei erwartungsgemäß großen Schwankungen (z.B. weil sich das Signal bereits in der vergangenen Zeit stark geändert hat) mehr Bits verwendet werden um die Schwankung genauer abbilden zu können.

Verallgemeinert man das Prinzip, so kommt man zur **Linear Predictive Coding** (LPC). Dabei wird jedes Sample als gewichtete Summe der n vorhergegangenen Samples zuzüglich einer mitübertragenen Differenz berechnet.

3.4.2 Parametrische Codierung

Parametrische Verfahren transformieren das Signal zuerst vom Zeit- in der Frequenzbereich. Dort werden dann Sprachsignalanalysen durchgeführt, um Parameter zu extrahieren, die das Signal möglichst gut beschreiben.

Grundlage dafür ist das Quelle-Filter-Modell (siehe Abbildung 2.8), dessen Modellparameter geschätzt werden.

Während mittels Signalformcodierungen (siehe 3.4.1) ohne hörbaren Qualitätsverlust nicht weniger als etwa 16 kBit/s erreicht werden kann, benötigen parametrische Verfahren üblicherweise 0,5-5 kBit/s. Sie sind aber natürlich auch ungleich aufwändiger und komplizierter. Außerdem erreichen sie nicht die Qualität von Signalformcodierungen. Als Vorteil ergibt sich aber zwangsläufig auch, dass sie sich nicht nur zur Sprachübertragung, sondern gleichsam zur Spracherkennung eignen, da dabei genau diese Parameter ausgenützt werden können.

Abbildung 3.1 zeigt das Grundprinzip. Auf Senderseite wird analysiert ob es sich um einen stimmhaften oder stimmlosen Laut handelt (**excitation analysis**: V/U), und im Falle eines stimmhaften Lautes, welche Sprachgrundfrequenz vorherrscht. In der **vocal tract analysis** wird die Einhüllende im Frequenzspektrum F als Approximation des menschlichen Vokaltraktes, sowie ein Amplitudenfaktor G ermittelt. Auf der Empfängerseite wird durch Simulation entsprechend dem Quelle-Filter-Modell das Signal anhand der Parameter rekonstruiert.

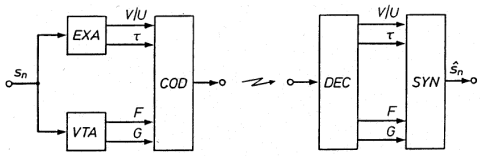

Abbildung 3.1: Parametrische Verfahren

Als Subtypen der parametrischen Verfahren lassen sich **Kanalvocoder** und **LPC-Vocoder** nennen. Sie unterscheiden sich darin, wie der Vokaltrakt approximiert wird: Kanalvocoder verwenden eine stückweise konstante Approximation, LPC-Vocoder eine genauere Kurve.

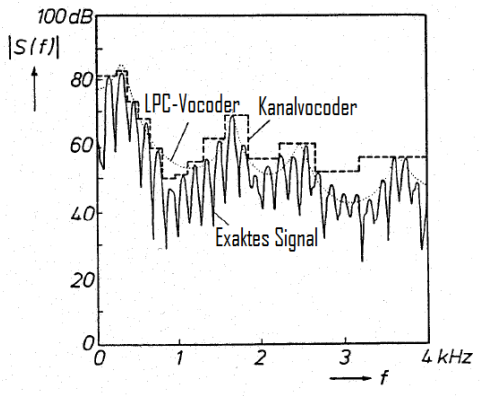

Abbildung 3.2: Vocodertypen

Bei der Vokaltraktanalyse eines Kanalvocoders (siehe Abbildung 3.3) werden in der Praxis einfach mehrere Bandpässe parallel geschaltet (gefolgt von einer Gleichrichtung und einem Tiefpass), die die Stärke der Anwesenheit bestimmter Frequenzbereiche im Gesamtsignal ermitteln. Auf der Empfängerseite (bzw. bei der Wiedergabe) wird die Anregung durch eine Schwingung samt Obertöne bzw. einen Noisegenerator (stimmhaft/stimmlos) simuliert. Anschließend wird der Vokaltrakt simuliert, indem die übertragenen

Filtereigenschaften auf das erzeugte Signal angewandt werden, d.h. gewisse Frequenzbereiche werden herausgefiltert und andere durchgeschleust.

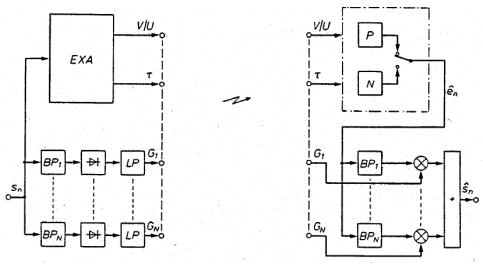

Abbildung 3.3: Kanalvocoder

Ein LPC-Vocoder übernimmt hingegen das Prinzip der LPC-Modulation im Falle der Signalformcodierung (linear predictive coding). Bei der gewöhnlichen LPC-Modulation wird ja, wie oben beschrieben, ein Signalwert durch Linearkombination der vergangenen Werte geschätzt, und schließlich nur noch die Differenz zum tatsächlichen Wert übertragen.

LPC-Vocoder (siehe Abbildung 3.4) funktionieren prinzipiell ähnlich, nur dass es sich dabei ja um ein parametrisches Verfahren statt um eine Signalformcodierung handelt. Die Filterfaktoren geben an, wie stark ein Frequenzband zu einem Zeitpunkt vertreten war. Zum Zeipunkt n wird nun versucht, diese Faktoren durch Linearkombination der vorhergehenden p Faktoren (a_1 bis a_p) zu schätzen. Die Gewichtungsfaktoren werden dabei so gewählt, dass die Abweichung vom tatsächlichen Vokaltrakt S_n minimal ist.

Was nun zu übertragen bleibt, sind nur noch die Gewichtungsfaktoren, da hier die eigentliche Information steckt. Das (geringe) Restsignal kann vernachlässigt werden (vgl. dazu: Hybride Verfahren, siehe Abschnitt 3.4.3). Dass sich diese Faktoren wesentlich seltener ändern als der Luftdruck bei direkter Signalformcodierung, führt letztendlich zur wesentlich geringeren Bitrate.

Parametrische Verfahren haben neben der geringen Bitrate den Vorteil, dass die Parameter relativ gut interpretiert werden können (im Vergleich zum Luftdruck bei der Signalformcodierung). So kann etwa die Signalgrundfrequenz ausgelesen oder verändert werden, es können Formantenlagen erkannt und bei Lautgrenzen interpoliert werden, usw.. Das ist auch die Grundlage für die Sprachsynthese und Spracherkennung.

Anmerkung: Kanalvocoder werden heute vor allem für die Erzeugung akustischer Spezialeffekte verwendet. Es erfolgt damit eine klare Trennung

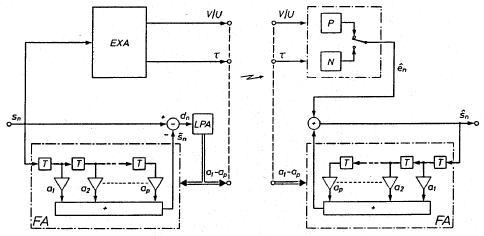

Abbildung 3.4: LPC-Vocoder

zwischen Lautgebung und Artikulation gemäß Quelle-Filter-Modell. Dadurch wird es einem ermöglicht, die Lautgebung gänzlich auszutauschen, die Artikulation aber beizubehalten. Dadurch kann man einem beliebigen Audiosignal, etwa einem Windgeräusch, einem Klavierspiel, usw., eine menschliche Sprache aufprägen. Das wäre damit zu vergleichen, dass bei einem menschlichen Sprecher die Simmbänder durch eine andere Signalquelle getauscht werden, der Mund-Nasen-Rachenraum aber gleich bleibt.

3.4.3 Hybride Verfahren

Aufgrund der Funktionsweise parametrischer Verfahren ist damit nie die Qualität einer Signalformcodierung erreichbar, selbst wenn die Bitrate erhöht wird. Abhilfe verschaffen hybride Verfahren, bei denen nun das in Abschnitt 3.4.2 unter LPC-Vocoder beschriebene Restsignal eben nicht verworfen wird. Stattdessen wird es mit geringer Bitrate auch noch übertragen.

3.4.4 Sprachsynthese

Bei der Sprachsynthese sind zwei voneinander unabhängige Teilschritte notwendig: zum einen die **Transkription** (Textumsetzung), und die eigentliche **Signalgenerierung**.

Transkription

Bei der Transkription wird der geschriebene Text für die Ausgabe aufbereitet, indem Buchstaben auf Lautfolgen abgebildet werden. Die Zuordnung ist hier alles andere als eindeutig, da Buchstaben unter Umständen sehr unterschiedlich ausgesprochen werden, je nachdem wo sie im Text vorkommen. Neben einer Folge von Lauten will man aber auch noch eine Prosodie für ganze Teilsätze und Sätze erzeugen, um eine möglichst natürliche Aussprache zu erreichen. Ausgabe dieses Teilschritts ist somit eine *Lautfolge* mit zusätzlichen *Markern für die Prosodie.*

Probleme sind dabei vor allem bei Mehrdeutigkeiten zu erwarten: Buchstaben werden nicht immer gleich ausgesprochen, Zahlen werden je nach Kontext mit unterschiedlichen Endungen gesprochen (eins, erster, erstens, usw.), Vor- und Nachsilben verschleiern den Stamm eines Wortes, zusammengesetzte Wörter, usw..

Es gibt folgende Strategien für eine gute Transkription. Allen Ansätzen ist gemeinsam, dass sie in irgendeiner Weise die Wortstruktur erfassen und Rückschlüsse auf die Aussprache ziehen wollen.

- **Regeln** für die Aussprache in einer Sprache, z.B. kurz gesprochene Vokale vor einem Doppelkonsonanten
- **Wörterbücher** (samt Lautschrift) für häufige Ausnahmen und Lehn- und Fremdwörter
- **Abtrennen von Prä- und Suffixen** um die Wortherkunft und somit die Aussprache herauszufinden
- **Trennen von zusammengesetzten Wörtern** und getrennte Aussprache für beide Teile (Clusteranalyse)

Um eine Prosodie für ganze Sätze zu entwickeln, wird eine (zumindest grobe) Syntaxanalyse des natürlichsprachlichen Satzes angestrebt. Heute versucht man üblicherweise, Wörterbücher für die wichtigsten Wörter (vor allem für Ausnahmen - die gerade bei häufigen Wörtern aufgrund der Natürlichkeit der Sprache vermehrt auftreten) zu entwerfen. Für nicht einhaltene Wörter greifen dann Regeln, die sozusagen eine gute Annäherung in unerwarteten Fällen darstellen.

Signalgenerierung

Es gibt grundsätzlich zwei Ansätze die durch die Transkription erzeugte Lautfolge auszugeben: die Laute werden tatsächlich entsprechend dem Quelle-Filter-Modell künstlich produziert (**artikulatorische Synthese**), oder es werden die einzelnen Laute von einem menschlichen Sprecher vorgesprochen und dann zusammengesetzt, wobei für Subphoneme mehrere Ausprägungen gespeichert sind. Bei der Synthese wird dann die Variante gewählt, so dass sie mit ihren Nachbarlauten möglichst gut zusammengefügt werden kann und glatte Übergänge entstehen: **Unit selection**.

Für die artikulatorische Synthese gab es einige Ansätze, bei denen durch genaue Analyse von Röntgenvideos Modelle des menschlichen Resonanzraumes gebaut wurden. Weitaus häufiger wird jedoch die Unit selection eingesetzt.

Wie leicht einsichtig ist, entstehen durch naives Zusammensetzen einzeln gesprochener Laute (im Deutschen ca. 30-40 verschiedene) hörbare Übergänge. Dass bloßes Zusammenfügen schlecht funktioniert, liegt vor allem daran, dass sich Laute nicht sprungartig ändern, sondern ineinander übergehen. Außerdem gibt es von ein und demselben Phonem in der Regel mehrere Abstufungen, die zwar nicht bedeutungsunterscheidend sind, jedoch das Natürliche in der Sprache ausmachen.

Diese unerwünschten Effekte gilt es zu minimieren. Es gibt dabei folgende Ansätze:

- Speichern der Lautübergänge statt der Laute (Diphonsynthese). Dadurch fallen die Schnittstellen zwischen zwei zusammengesetzten Fragmenten nicht auf die Übergänge, sondern auf die relativ gleichmäßigen Mitten der Laute, weshalb sie weitaus weniger hörbar sind.
- Aufnahme mehrerer kontextabhängiger Lautabstufungen (Allphonsynthese)
- Aufnahme größerer Einheiten (bis hin zu Silben und ganzen Wörtern), zumindest für häufige Phrasen. Da die 100 häufigsten Wörter im Deutschen schon 50% des Textes ausmachen, kann man mit relativ geringem Aufwand starke Verbesserungen erzielen.

Mit algorithmischen Verfahren (bei parametrischer Signalspeicherung) kann man die Übergänge (die man davor idealerweise mit obigen Mitteln minimiert) noch weiter glätten. Heute verwendet man üblicherweise eine Kombination aus Unit Selection und Glättung.

Kapitel 4

Spracheingabe

Es lässt sich grundsätzlich zwischen **Sprechererkennung** und **Spracherkennung** unterscheiden. Beides untergliedert sich in 3 Teilschritte: Sprachsignalvorverarbeitung, Merkmalextraktion und Klassifikation.

4.1 Sprechererkennung

Aufgabe der Sprecherekennung ist es, den Urheber eines gesprochenen Textes zu erkennen, während der Textinhalt irrelevant ist, da dieser bekannt ist. Notfalls kann der Text auch von einem Menschen eingegeben werden.

Es lässt sich weiter unterscheiden ob ein Sprecher **verifiziert** werden soll (ob jemand eine bestimmte Person ist: ja/nein, z.B. bei Sicherheitssystemen), oder ob ein Sprecher **identifiziert** werden soll (mit welcher Wahrscheinlichkeit ein Text von einer aus mehreren Personen stammt, z.B. in der Kriminalistik).

Die Grundidee ist es, den *textabhängigen* Teil des Signals herauszufiltern, und nur den *sprecherabhängigen* Teil zu erhalten. Dieser kann dann mit einem gespeicherten Sprecherprofil verglichen werden.

Das Signal wird dazu wieder in den Frequenzbereich übertragen und dort über eine gewisse Zeit beobachtet, also ein Sonagramm erstellt. Über die Zeit wird dort ein „Durchschnittswert“ (Details folgen) ermittelt, um den textabhängigen Teil herauszufiltern. Was bleibt, ist somit eine spektrale Energieverteilung (der sogenannte Merkmalsvektor), die für einen Sprecher charakteristisch ist.

„Durchschnittsbildung“ funktioniert dabei mit dem arithmetischen Mittel natürlich nur, wenn beide Sprachproben entweder denselben oder genügend lange Texte umfassen (damit sich die Unterschiede statistisch aufheben). Um

auch kurze unterschiedliche Texte vergleichen zu können, kann ein gewichteter Durchschnitt verwendet werden. Je nachdem, wie häufig die einzelnen Laute im Text vorkommen, sind entsprechend der Formantenlagen dieser, bestimmte Energieverteilungen zu erwarten.

In diesem Zusammenhang sind die **False Acceptance Rate** (FAA) und die **False Rejection Rate** (FRA) zu erwähnen. Sie geben jeweils an, wie viele berechtigte User fälschlicherweise abgewiesen werden, bzw. wie viele Unberechtigte fälschlicherweise Zutritt bekommen. Je nach Anwendung kann das eine auf Kosten des anderen verbessert werden, oft wird jedoch eine gute Balance angestrebt (auch um Systeme vergleichbar zu machen).

4.2 Spacherkennung

Bei der Spracherkennung lassen sich folgende grundsätzliche Zielsetzungen unterscheiden:

- Die Sprache soll **interpretiert** werden und einen Prozess auslösen („command and control“), z.B. Sprachwahl beim Handy. Dabei ist der Sprecher meist bekannt. Hier kommen oft nur wenige unterschiedliche Wörter vor.
- Die Sprache soll **erkannt** und in Schrift umgewandelt werden. Solche Systeme sind üblicherweise (noch) sprecherabhängig, weil man sie am Beginn trainieren muss.
- Sie soll semantisch **verstanden** und beantwortet werden. Einsatzbereiche sind zum Beispiel im automatisierten Kundenservice zu finden, weshalb solche Systeme sprecherunabhängig sein müssen.

Man bezeichnet diese Systeme auch als: **Einzelworterkennung**, **Diktiersystem** und **Dialogsystem**.

Das Hauptproblem bei der Spracherkennung ist, dass es zu einem Satz sehr viele verschiedene akustische Realisierungen gibt (Phon - Phonem). Das trifft sowohl innerhalb eines Sprechers, und erst recht bei verschiedenen Sprechern zu (Intra- und Inter-Sprecher-Variabilität). Auch das Feststellen der Wortgrenzen ist nicht trivial (es gibt im Allgemeinen keine Pausen zwischen Wörtern!). Vor allem die Sprechgeschwindigkeit variiert sehr stark.

Weitere Probleme betreffen das Auftreten ungewöhnlicher Wörter (d.h. welche, die nicht im Wörterbuch sind), menschliche Fehler beim Sprechen (Spontansprache), Umgebungsgeräusche und Signalübertragungsfehler.

Umgekehrt wie bei der Sprechererkennung, gilt es nun die sprecherabhängigen Merkmale zu eliminieren und nur die sprachabhängigen zu erhalten.

4.2.1 Funktionsweise im Detail

Die 3 Teilschritte bei der Spracherkennung sind die Signalvorverarbeitung, der Mustervergleich und die Klassifizierung.

Signalvorverarbeitung

Die Signalvorverarbeitung verfolgt folgende Ziele:

- Datenreduktion
- Parameterextraktion
- Gewinnung von Merkmalen
- Störgeräusche herausfiltern
- Wortgrenzen erkennen

Die Merkmalsextraktion hängt davon ab, wie der später folgende Vergleich aufgebaut ist. Üblich sind Energieniveaus in einzelnen Frequenzbänden, Nulldurchgänge und *Mel Frequency Cepstral Coefficients (MFCC).*

Bei MFCC wird das Signal in den Frequenzbereich transformiert, dort logarithmiert, und wieder in den Zeitbereich rücktransformiert. Dieser logarithmische Ansatz entspricht auch dem menschlichen Hörempfinden.

Meistens ist die Signalvorverarbeitung (unter anderem) eine Transformation des Signals in den Frequenzbereich. Dazu kann man zum Beispiel die *Fast Fourier Transformation* (FFT) verwenden. Daneben können noch Nulldurchgänge, Autokorrelationskoeffizienten, die Sprachgrundfrequenz und ähnliches als Parameter gewonnen werden.

Mustervergleich

Der Vergleich zwischen einem eingegebenen Sprachsignal und den in Frage kommenden gespeicherten Referenzmustern ergibt numerische Abstände, die im folgenden Schritt für die Klassifizierung ausgewertet werden. Zunächst gilt es also den Abstand zwischen 2 Sprachproblem, d.h. ihrer Merkmalsmatrizen (=Merkmalsvektoren im Zeitverlauf), zu bestimmen.

Das Hauptproblem beim Vergleich ist der zeitliche Verlauf. Die Sprechgeschwindigkeit variiert sehr stark. Diese Skalierung variiert selbst ebenfalls

im Zeitverlauf (weshalb eine einfache Streckung oder Stauchung nicht in Frage kommt).

Dynamic Time Warp (DTW), *Hidden Markov Models* (HMM) und Neurale Netze sind übliche Lösungsansätze, wobei sich letzteres nicht durchgesetzt hat.

DTW erzeugt für die beiden zu vergleichenden Muster ein Kurzzeitsonagramm. Dieses gibt bekanntlich für einzelne Frequenzbänder an, wie viel akustische Energie dort vorhanden ist. Es wird nun ein zweidimensionales Diagramm erzeugt, wobei die Dimensionen den beiden Mustern entsprechen, und an einem Schnittpunkt aufgetragen wird, wie ähnlich die beiden Energievektoren zum jeweiligen Zeitpunkt sind (siehe Abbildung 4.1). Bei ähnlichen Wörtern zeichnet sich im Wesentlichen ein Weg von links unten nach rechts oben ab (der jedoch nicht immer genau antlang der Diagonalen verläuft), während bei unterschiedlichen Wörtern kein solcher Weg existiert.

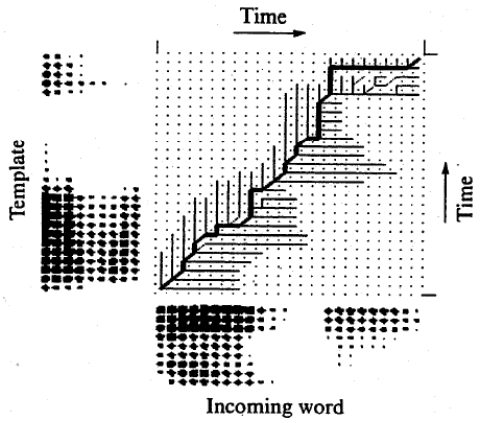

Abbildung 4.1: Mustervergleich

Gesucht wird also ein optimaler Weg von links unten nach rechts oben. Dazu wird dynamische Programmierung eingesetzt. Es werden dabei die lokalen Kosten für jedes Feld berechnet, wobei große Unterschiede hohe Kosten und kleine Unterschiede niedrige Kosten verursachen. Anschließend wird spaltenweise von links nach rechts vorgegangen, und für jedes Feld einer Spalte (von unten nach oben) der globale optimale Weg wie folgt berechnet: es kann davon ausgegangen werden, dass der kürzeste Weg für alle Felder in der Spalte links von der aktuellen, bereits berechnet wurde. Somit kann man durch simples Addieren der Kosten zu allen potentiellen Vorgängerpunkten (deren Wege wie gesagt sicher schon optimal sind) den optimalen Weg bestimmen.

Dieser Algorithmus funktioniert hier nur deshalb, weil die Zeit nicht rückwärts laufen kann und ist kein allgemeiner Wegsuchalgorithmus.

Um im Vergleich zur vollständigen Enumeration einen bedeutenden Performancevorteil zu erzielen, werden weitere Prinzipien aus der Optimierungstheorie angewendet. Spezifisch für die Sprachverarbeitung kann man den Suchraum für den optimalen Weg weiter einschränken, indem man gewisse Eigenschaften des Diagramms ausnutzt, etwa dass die Funktion monoton und die Steigung beschränkt ist.

Vorzeitiges Abbrechen ist vor allem bei der Einzelworterkennung von Vorteil, wenn der Benutzer dazwischen auch „normal“ reden darf, also keine Steuerkommandos eingeben will. In diesem Fall versucht das System laufend anhand des Ähnlichkeitsgraphen einen Weg zu finden. Nur wenn ein Weg bis zum Ende gefunden wird, wird das Wort als Steuerwort erkannt.

Heute werden jedoch überwiegend statistische Verfahren eingesetzt: *Hidden Markov Models* (HMM). Statistische Modelle beschreiben die möglichen unterschiedlichen Realisierungen einer Erkennungseinheit, z.B. eines Wortes. Die Modelle geben die Wahrscheinlichkeiten dafür an, dass diese Einheit einer Folge von akustischen Einheiten entspricht.

Das **Spracherkennungproblem** lautet also: Es ist eine Folge a von akustischen Eingaben gegeben. Finde die wahrscheinlichste Folge von Wörtern. Mathematisch beschrieben: $ArgMax_w(P(a\|w)*P(w)/P(a)) = ArgMax_w(P(a\|w)*P(w))$. $P(w)$ gibt die Auftrittswahrscheinlichkeit eines Wortes an (aus dem *Sprachmodell*), $P(a\|w)$ die Wahrscheinlichkeit für die akustische Folge a unter der Annahme, dass es tatsächlich w war. $P(a\|w)$ stammt aus dem *akustischen Modell*, das damit die unterschiedlichen Realisierungen abdeckt.

Tatsächlich implementiert werden die Prinzipien als *Hidden Markiv Model*. *Markov Models* sind Graphen, deren Knoten Zustände n_i repräsentieren und deren Übergänge mit Wahrscheinlichkeiten beschriftet sind.

Hidden Markov Models sind eine Erweiterung, die einen doppelten stochastischen Prozess beschreiben: zum einen aus Zuständen (auch innere Zustände genannt - Erklärung folgt) und deren Übergänge, zum anderen die möglichen unterschiedlichen Merkmalsvektoren innerhalb eines Zustandes. Innerhalb der Zustände ist somit festgelegt, mit welcher Wahrscheinlichkeit (die sogenannten Emissionswahrscheinlichkeiten) bestimmte Merkmalsvektoren auftreten. Den Zuständen an sich ist keine bestimmte Bedeutung zuordenbar (daher „hidden“), sie werden vom Trainingsalgorithmus festgelegt um eine gewisse Zielfunktion zu optimieren.

Pro sprachlicher Einheit (ein Wort, eine Silbe, etc.) wird ein HMM aufgebaut. Anschließend wird es verwendet um für eine Eingabe eine Wahrscheinlichkeit zu berechnen: für jedes Modell wird statistisch berechnet, wie groß die Wahrscheinlichkeit ist, dass dieses Modell die eingegebene Folge von Merkmalen erzeugt.

Beim Aufbau der Modelle müssen umgekehrt die Zustandsfolge und die Modellparameter so gewählt werden, dass die Wahrscheinlichkeit für die Trainingsdaten maximiert wird.

Bei der Verarbeitung von Sprache kann für die HMMs die vereinfachende Annahme getroffen werden, dass nie in einen zuvor bereits verlassenen Zustand zurückgewechselt wird (Links-Rechts-Modelle), oder dass eine maximale Sprunggröße beim Wechseln in Foglezustände nicht überschritten werden darf.

Anmerkung: Die Anzahl der Zustände ist meistens für ein konkretes System fixiert, könnte aber grundsätzlich für jede sprachlich Einheit auch anders sein.

Es bleiben nun also folgende Aufgaben zu lösen:

- **Training**:
  (i) Welche Zustandsfolge ist in einem Modell zu wählen, damit das Sprachsample möglichst gut beschrieben wird?
  (ii) Welche Modellparameter beschreiben das eingegebene Sprachsample am besten?

- **Spracherkennung**:
  Wie hoch ist die Wahrscheinlichkeit, dass die Spracheingabe von einem bestimmten Modell stammt?

Für alle Teilaufgaben gibt es komplexe Algorithmen. Eine besondere Stärke der HMMs ist, dass es Algorithmen zur Anpassung eines gut trainierten Systems an einen neuen Sprecher gibt.

Arten von Fehlern sind falsche Worteinfügungen oder -auslassungen, oder Wortverwechslungen (das gleiche gibt es auch auf Lautebene).

4.2.2 Erweiterungen

Obige Prinzipien sind für Einzelworterkennung gedacht. Sie können wie folgt auf andere Anwendungen erweitert werden.

Sprecherunabhängigkeit kann über Mittelung oder Frequenznormalisierung der Merkmalsvektoren erreicht werden. Alternativ dazu können auch für unterschiedliche Sprecher einfach mehrere HMMs erstellt werden.

Größeres Vokabular kann durch kleinere Erkennungseinheiten, z.B. von Lauten statt ganzen Wörtern, erreicht werden.

Nicht nur einzelne Wörter, sondern ganze Sätze zu erkennen, ist nicht trivial. Die einfachste Variante davon nennt man *wort spotting* und meint, dass in der gesprochenen Sprache lediglich Einzelwörter erkannt und herausgefiltert werden sollen.

Verbessert werden kann die Erkennungsrate durch statistische Modelle für die Abfolge von Wörtern.

4.2.3 Training von Modellen

Es sind große Datenmengen von 100 - 500 Stunden Sprachmaterial notwendig, um sprecherunabhängige Systeme zu trainieren. Diese Daten müssen nicht nur gesprochen, sondern müssen auch noch in Lautschrift übersetzt werden. Die ausgewählten Sprecher sollten natürlich repräsentativ für das Zielpublikum sein (z.B. Alter, Geschlecht).

Wenn notwendig, dann kann man auch mehrere Sprecher einsetzen, oder das System an den den konkreten Enduser im Rahmen eines Kurztrainings noch anpassen (sprecheradaptive Systeme). Selbstkorrigierende Systeme sind auch denkbar, die ihre Modelle bei Korrekturen des Benutzers über die Tastatur anpassen.

Kapitel 5

Sprachdialogsysteme

5.1 Einleitung

Sprachdialogsysteme sind eine Kombination aus Spracherkennung und Sprachausgabe, mit dazugeschalteter Dialogsteuerung. Die Steuerungskomponente greift schließlich auf ein Anwendungssystem zu.

Beispiele: Auskunftssysteme, Flugbuchung, Vorlesen von E-Mails, *Voice Browsing*.

Vorteile solcher Systeme sind Kosteneinsparungen, kürzere Wartezeiten für Kunden, neuartige Anwendungen.

5.2 Dialoginitiative

Die Steuerung des Dialoges kann auf mehrere Arten geschehen, abhängig davon, ob der menschliche Benutzer oder das System die Initiative übernimmt.

Bei unregelmäßig benutzten Systemen ist in der Regel eine **Systeminitiative** erwünscht, die den Benutzer durch den Dialog führt. Erfahrene Benutzer bevorzugen normalerweise dagegen eher die **Benutzerinitiative**, bei der sie unabhängig vom System ihre Kommandos eingeben können. Ein Kompromiss der beiden Varianten ist die **gemischte Initiative**, bei der der Benutzer grundsätzlich frei Kommandos eingeben kann, bei Bedarf aber eine Art Hilfefunktion aufrufen kann, die den Benutzer dann wieder führt.

Oft erwünscht ist auch ein Mechanismus, den man als *Barge-In* bezeichnet. Dabei kann der Benutzer (vor allem bei längeren Antworten) das System unterbrechen. Die Steuerungskomponente muss dann natürlich feststellen, an welcher Stelle im Dialog fortgesetzt werden soll. Außerdem gilt es hier Störgeräusche unbedingt als solche zu erkennen, damit der Dialog

nicht fälschlicherweise abgebrochen wird.

Erkennungsfehler sollten natürlich so schnell wie möglich erkannt werden, um im Dialog nicht zu weit zurückspringen zu müssen. Die **Verifikation** der Benutzereingaben kann dabei **explizit** erfolgen, indem ausdrücklich noch einmal nachgefragt wird (Ja/Nein-Frage, da diese Wörter leicht erkannt werden können), oder man baut die Überprüfung versteckt in die nächste Frage ein (**implizit**).

5.3 Dialogmodellierung

Dialoge können auf folgende Arten modelliert werden:

- Als Ablaufdiagramm, bei dem ein strikter Ablauf des Dialogs vorgegeben werden. Freilich ist diese Variante nur für einfache Systeme geeignet.
- *slot-filling* ist ein Konzept, bei dem lediglich bestimmte Informationseinheiten wie Name und Adresse erfragt werden, die Reihenfolge aber irrelevant ist. Das System fragt nur dort nach, wo noch Antworten offen sind.
- Planbasierter Ablauf ist ein Prozess, bei dem man nur ein Ziel festsetzt. Das System versucht dann selbstständig einen entsprechenden Dialog zusammenzustellen.

Eine konkrete Technologie zum Beschreiben solcher Systeme ist VoiceXML, das nach dem Prinzip des slot-filling arbeitet. Es ist ein Quasi-Standard und akustisches Pendant zu XHTML. Es können darin „akustische Eingabefelder“ definiert werden. Dazu legt man fest, was das System als Label ausgeben soll (Prompt), und welche Eingaben in diesem Schritt erlaubt sind (sogenannte Grammatiken, meist kontextfreie).

Literaturverzeichnis

[1] Dr. Markus Kommenda - TU Wien
Vorlesung und begleitende Folien - Wintersemester 2008/09

[2] Deutsche Wikipedia - 30.11.2008

[3] Uwe Reichel - Lecture Notes
http://www.phonetik.uni-muenchen.de/
∼reichelu/kurse/sonagramme/sona_slides_2.pdf

[4] http://www.ling.uni-potsdam.de/∼mayer/teaching/phonetik/beispielspec.pdf

[5] Jochen Trommer - Lecture Notes
http://www.uni-leipzig.de/
∼jtrommer/phonetik07/k6a.pdf