(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Veröffentlichungsnummer: **0 157 903**
B1

(12) # EUROPÄISCHE PATENTSCHRIFT

(45) Veröffentlichungstag der Patentschrift: **13.01.88**

(51) Int. Cl.⁴: **G 10 L 5/04**

(21) Anmeldenummer: **84109492.3**

(22) Anmeldetag: **09.08.84**

(54) **Verfahren und Anordnung für die Sprachsynthese.**

(30) Priorität: **23.02.84 DE 3406540**

(43) Veröffentlichungstag der Anmeldung: **16.10.85 Patentblatt 85/42**

(45) Bekanntmachung des Hinweises auf die Patenterteilung: **13.01.88 Patentblatt 88/2**

(84) Benannte Vertragsstaaten: **FR GB IT NL**

(56) Entgegenhaltungen:
EP-A-0 059 832
EP-A-0 114 123

ACUSTICA, Band 26, 1972, Seiten 33-36, Stuttgart, DE; W. ENDRES: "The transitional sounds of the German language as link elements for a speech synthesis"
IEEE COMMUNICATIONS MAGAZINE, Band 21, Nr. 9, Dezember 1983, Seiten 26-34, IEEE, New York, US; D. O'SHAUGHNESSY: "Automatic speech synthesis"
ICASSP 80 - PROCEEDINGS-IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SINGAL PROCESSING, 9.-11. April 1981, Denver, US, Seiten 557-563, IEEE, New York, US; S. IMAI u.a.: "Cepstral synthesis of Japanese from CV syllable parameters"

(73) Patentinhaber: **Matth. Hohner AG, Hohnerstrasse 8, D-7218 Trossingen (DE)**

(72) Erfinder: **Deforeit, Christian, 6, rue de l'Hôpital, F-21140 Semur- en- Auxois (FR)**

(74) Vertreter: **Dipl.- Ing. H. Marsch Dipl.- Ing. K. Sparing Dipl.- Phys.Dr. W.H. Röhl Patentanwälte, Rethelstrasse 123, D-4000 Düsseldorf (DE)**

EP 0 157 903 B1

Anmerkung: Innerhalb von neun Monaten nach der Bekanntmachung des Hinweises auf die Erteilung des europäischen Patents im Europäischen Patentblatt kann jedermann beim Europäischen Patentamt gegen das erteilte europäische Patent Einspruch einlegen. Der Einspruch ist schriftlich einzureichen und zu begründen. Er gilt erst als eingelegt, wenn die Einspruchsgebühr entrichtet worden ist (Art. 99(1) Europäisches Patentübereinkommen).

LIBER, STOCKHOLM 1988

Beschreibung

Die Erfindung betrifft ein Verfahren für die Sprachsynthese und eine Anordnung zu seiner Durchführung.

Es ist bekannt, die in einer Sprache vorkommenden Phoneme (Konsonanten und Vokale) einzeln abzuspeichern und dann bedarfsweise sequentiell auszulesen. Da die Anzahl der Phoneme relativ gering ist, bietet sich dabei die Möglichkeit, mit relativ wenig Speicherkapazität zu arbeiten. Nachteilig ist jedoch, daß bei der Wiedergabe ein von natürlicher Sprache stark abweichender Klang entsteht, weil zwischen aufeinanderfolgenden Konsonanten und Vokalen ein "Klick-"geräusch hörbar wird, so daß dieses Prinzip praktisch nicht angewandt wird.

Um eine der menschlichen Sprache weitgehend angenäherte synthetische Sprache zu erzeugen, geht man daher so vor, daß jeweils aus den vorkommenden Konsonanten und den vorkommenden Vokalen gebildete Kombinationen, sogenannte Di-Phoneme, gespeichert und ausgelesen werden. Es versteht sich, daß hierfür eine sehr erhebliche Speicherkapazität benötigt wird, die für eine indogermanische Sprache in der Größenordnung von etwa 500 oder mehr Di-Phonemen liegt. Die Verwendung des Verfahrens beschränkt sich demgemäß auf solche Fälle, bei denen ein erheblicher Aufwand gerechtfertigt ist.

Ausgehend von dem letztgenannten Verfahren liegt der Erfindung die Aufgabe zugrunde, den Speicheraufwand erheblich herabzusetzen, so daß das Verfahren auch bei Massenkonsumgütern, z. B. Spielzeugen, (Puppenstimmen und dergleichen) anwendbar wird.

Der Patentanspruch 1 definiert das erfindungsgemäße Verfahren. Man erkennt, daß jeder Konsonant nur in Kombination mit einem einzigen Einheitsvokal abgespeichert wird, der hier und im folgenden mit "&" bezeichnet werden soll, und der etwa dem "e" im deutschen Wort "alle", im englischen Wort "the", im französischen Wort "le" entspricht. Es hat sich gezeigt, daß bei entsprechend zeitversetztem Auslesen das & von dem dann wiedergegebenen eigentlichen Vokal völlig oder jedenfalls soweit maskiert wird, daß die resultierende Silbe der natürlichen Aussprache weitgehend nahekommt.

Bei der üblichen digitalen Abspeicherung der Phoneme ist es zweckmäßig, bei den Konsonanten an der richtigen Stelle einen Befehl abzuspeichern, der den Beginn der Auslesung des Vokalspeichers einleitet. Ferner kann es zweckmäßig sein, durch entsprechend ausgelegte Filter die Frequenzen der Konsonanten einerseits, der Vokale andererseits unterschiedlich zu verstärken bzw. zu bedämpfen, um die Maskierung der & zu verbessern.

Der Patentanspruch 5 definiert eine Anordnung gemäß der Erfindung, mit der die Sprachsynthese durchführbar ist.

Unter Bezugnahme auf die beigefügten Zeichnungen soll die Erfindung nachstehend im einzelnen erläutert werden.

Fig. 1 zeigt anhand eines Beispiels das Prinzip des Verfahrens,

Fig. 2 zeigt Frequenzgänge von Filtern für Vokale und Konsonanten,

Fig. 3 zeigt schematisch ein mögliches Speicherformat für Konsonanten und Vokale,

Fig. 4 zeigt eine bevorzugte Hüllkurve für die Konsonantenerzeugung,

Fig. 5 ist ein Blockdiagramm einer Anordnung zur Ausführung des Verfahrens, und

Fig. 6 ist ein Diagramm zur Darstellung des Zeitablaufs bei der Synthese eines einfachen Wortes.

Da der Auslesevorgang zeitversetzt, däs heißt so erfolgt, daß die Auslesung eines Vokals bereits beginnt, während noch das Auslesen des Konsonanten-Di-Phonems (nämlich dessen &-Teil) abläuft, arbeitet man mit zwei Auslesekanälen. In Fig. 1 stellt das obere Diagramm den Hüllkurvenverlauf des Konsonantenkanals, das untere den des Vokalkanals dar, wobei als Beispiel das einfache Wort "DATO" gewählt ist. Man erkennt, daß gleichzeitig die &-Anteile des Konsonantenkanals und die Vokale wiedergegeben werden, und bereits dadurch werden die schwachen &-Laute stark maskiert. Diese Maskierung kann aber noch durch weitere Maßnahmen unterstützt werden.

Fig. 2 stellt ein erstes Mittel hierfür dar. Es ist bekannt, daß das Frequenzspektrum der Konsonanten und Vokale unterschiedlich ist; z. B. liegen bei einer männlichen Stimme die Maxima der Konsonanten im Bereich von etwa 600...3000 Hz, der Vokale im Bereich von etwa 200...1000 Hz. Dementsprechend werden den beiden Kanälen Filter mit den in Fig. 2 gezeigten Durchlaßbändern zugeordnet, wobei die Filterung entweder bei der Aufzeichnung oder bei der Wiedergabe erfolgen kann.

Fig. 3 zeigt schematisch das Format für die Speicherung. Bei der Aufzeichnung werden die Laute digitalisiert, das heißt mit einem Takt von z. B. 10 KHz oder mehr amplitudenabgetastet und die so erhaltenen Daten werden in aufeinanderfolgenden Speicherplätzen für serielles Auslesen abgespeichert. Es werden jedoch zwei Speicherplätze für Kommandodaten freigehalten, nämlich ein Kommando "weiter" und ein Kommando "Ende". Das Kommando "weiter" bedeutet den Zeitpunkt, bei welchem der jeweils andere Kanal mit dem Auslesen fortfahren soll; dieses Kommando liegt bei den Konsonantendaten beim Übergang des eigentlichen Konsonantenlauts zum &-Teil, während es bei den Vokaldaten nahe dem Ende des Datenstrangs liegt. Das Kommando "Ende" versteht sich von selbst, ist aber erforderlich, weil die einzelnen Phoneme unterschiedliche Dauer besitzen. Das Kommando "weiter" kann dazu verwendet werden, um den Maskierungseffekt noch zu verstärken, indem bei

seinem Auftreten die Hüllkurve des gerade ausgelesenen Kanals bedämpft wird, wie in Fig. 4 angedeutet, wofür man ein übliches analog arbeitendes Dämpfungsglied aus Diode, Widerstand und Kondensator verwenden kann.

Fig. 5 zeigt in Blockform ein Ausführungsbeispiel eines Sprachsynthesizers, der - wie man erkennt - höchst einfach aufgebaut ist. Die Auswahl der wiederzugebenden Phoneme erfolgt durch externe Mittel, beispielsweise einen Mikroprozessor, und bildet keinen Gegenstand der vorliegenden Erfindung; hier ist deshalb nur als Block 1 eine externe Steuerschaltung angedeutet.

Die Anordnung umfaßt zwei untereinander identische Kanäle, von denen nachstehend nur einer beschrieben wird.

Ein Speicheradressenzähler 2 wird von der Steuerschaltung 1 auf eine bestimmte Phonem-Startadresse gesetzt. Ein Phonemspeicher 3 enthält alle für eine gegebene Sprache benötigten Phoneme, wobei für viele Sprachen sechsunddreißig Phoneme ausreichend sind. Die Phoneme sind nach Filterung bei der Aufnahme (wie oben erläutert) digitalisiert und in dem in Fig. 3 dargestellten Format abgespeichert; dabei können beispielsweise die Kodes "0" bzw. "1" für die Kommandos "weiter" bzw. "Ende" reserviert sein. Ein Taktgenerator 5 erzeugt den Auslesetakt von z. B. 10 KHz, und zwar für beide Kanäle. Die ausgelesenen Daten gelangen zu einem Dekoder 4, der feststellt, ob es sich um Daten oder eines der Kommandos "weiter" bzw. "Ende" handelt. Daten gelangen über einen Digital-Analog-Umsetzer 6 sowie ein Multiplizierglied 7 zu einem Summierglied 8 und von dort zu einer Verstärker-Lautsprecher-Einheit 9.

Bei Dekodierung des Kommandos "Ende" wird über ein UND-Gatter 10 die Inkrementierung des Adresszählers 2 gesperrt.

Wird das Kommando "weiter" dekodiert, so wird ein Phonem-Anforderungs-Flipflop 11 für den jeweils anderen Kanal gesetzt; seine Rücksetzung erfolgt durch die externe Steuerschaltung bei Eingabe der nächsten Startadresse. Ferner wird beim Kommando "weiter" ein Dämpfungsflipflop 12 umgeschaltet, das mit seinem Ausgang F dem einen, mit seinem Ausgang $\bar{F}$ dem anderen Kanal einen Hüllkurvengenerator 13 zuschaltet, der auf das Multiplizierglied 7 einwirkt, so daß der Ausgang des betreffenden Kanals sanft abfallend bedämpft wird, ohne daß jedoch das "Klick"-geräusch entsteht. Die Ausgänge beider Kanäle werden im Summierglied 8 kombiniert.

Der jeweilige Setzzustand des Flipflops 12 wird auch zu der externen Steuerschaltung übertragen, um dieser zu signalisieren, welcher der beiden Kanäle belegt werden kann, etwa zu Beginn eines Auslesezyklus nach Inbetriebnahme der Schaltung.

Bevor unter Bezugnahme auf Fig. 6 ein Synthesevorgang im einzelnen erläutert wird, sei noch auf mögliche Abwandlungen der in Fig. 5 gezeigten Blockschaltung hingewiesen.

Der Speicheraufwand läßt sich halbieren, wenn für beide Kanäle nur ein Phonemspeicher 3 vorgesehen ist und das Auslesen im Zeitmultiplex erfolgt. Das Multiplizierglied 7 ist in bestimmten handelsüblichen Digital-Analog-Umsetzern bereits enthalten, so daß man den Ausgang der Hüllkurvengeneratoren 13 nur mit dem entsprechenden Eingang des Umsetzers zu verbinden braucht. Man kann die Schaltung auch weitgehend in einem Mikroprozessor realisieren, wobei dann entweder die beiden Hüllkurvengeneratoren und die beiden Umsetzer außerhalb bleiben oder nur ein einzelner, gemeinsamer Umsetzer, während alle anderen Vorgänge vom Mikroprozessor digital durchgeführt werden.

Fig. 6 zeigt

- in Zeile (a) den Takt des Taktgenerators 5; dieser Takt kann starr sein, kann aber auch von der Steuerschaltung 1 variiert werden, um eine der natürlichen Sprache noch ähnlichere Phrasierung zu erzielen,
- in Zeile (b) Formate aus dem ersten Kanal, hier die Phoneme "D&" und "T&",
- in Zeile (c) den Logikpegel am Ausgang des Flipflop 12,
- in Zeile (d) den Logikpegel am Ausgang des Flipflops 11 des zweiten Kanals,
- in Zeile (e) Formate aus demselben zweiten Kanal, hier Phoneme "a" und "o",
- in Zeile (f) den Logikpegel am Ausgang des Flipflops 11 des ersten Kanals,
- in Zeilen (g) bzw. (h) die Hüllkurven, erzeugt von den Hüllkurvengeneratoren 13 des ersten bzw. des zweiten Kanals, und
- in Zeilen (i) bzw. (k) die analogen Ausgangssignale des ersten bzw. zweiten Kanals; dabei sind die Hüllkurven nicht als repräsentativ für die tatsächlich erzeugten Laute "D", "A", "T" oder "O" zu verstehen; das Diagramm dient nur der Erläuterung des zeitlichen Ablaufs.

Patentansprüche

1. Verfahren für die Sprachsynthese, bei dem Kombinationen aus je einem Konsonanten und einem ihm folgenden Vokal abgespeichert und bedarfsweise ausgelesen werden, dadurch gekennzeichnet, daß
   - jeder in der Sprache vorkommende Konsonant, zusammen mit einem schwachen Einheitsvokal "&" abgespeichert wird,
   - alle in der Sprache vorkommenden Vokale einzeln abgespeichert werden,
   - die gewünschte Konsonanten-Vokal-Kombination durch zeitversetztes Auslesen von Konsonant und Vokal in zwei Kanälen unter Maskierung des Einheitsvokals "&" durch den ausgelesenen Vokal gebildet wird.

2. Verfahren nach Anspruch 1, dadurch gekennzeichnet, daß beim Abspeichern der Kombinationen aus Konsonant und Einheitsvokal vokaltypische Frequenzen bedämpft werden und

daß beim Abspeichern der Vokale konsonantentypische Frequenzen bedämpft werden.

3. Verfahren nach Anspruch 1 oder 2, dadurch gekennzeichnet, daß die Amplituden der Einheitsvokale "&" beim Auslesen bedämpft werden.

4. Verfahren nach Anspruch 1, 2 oder 3, dadurch gekennzeichnet, daß das Auslesen mit variabler Taktfrequenz erfolgt.

5. Anordnung zur Sprachsynthese, bestehend aus einer Speichervorrichtung zur Speicherung von Konsonanten und Vokalen und einer Ausleseschaltung zum Auslesen von Kombinationen aus je einem Konsonanten und einem ihm folgenden Vokal, dadurch gekennzeichnet, daß

-jeder in der Sprache vorkommende Konsonant zusammen mit einem schwachen Einheitsvokal "&" in der Speichervorrichtung abgespeichert wird,

-alle in der Sprache vorkommenden Vokale einzeln in der Speichervorrichtung abgespeichert werden,

-die Ausleseschaltung aus zwei alternierend aktivierbaren Auslesekanälen und einem durch ein in einem Kanal ausgelesenes Kommando zur Aktivierung des anderen Kanals ansteuerbaren Umschaltkreis besteht, und die gewünschte Konsonanten-Vokal-Kombination aus den zwei Kanälen zeitversetzt ausliest, und daß

- für jeden Kanal ein vom Umschaltkommando aktivierbarer Hüllkurvengenerator vorgesehen ist, der eine Amplitudenbedämpfung zur Maskierung des Einheitsvokals "&" durch den ausgelesenen Vokal bewirkt.

6. Anordnung nach Anspruch 5, dadurch gekennzeichnet, daß jeder Kanal einen Speicher für alle benötigten Laute (Phoneme und Di-Phoneme) umfaßt.

7. Anordnung nach Anspruch 5 oder 6, bei der die Laute (Phoneme und Di-Phoneme) digital in jedem Speicher abgespeichert und sequentiell auslesbar sind, dadurch gekennzeichnet, daß zumindest bei den Di-Phonemen das Umsteuer-Kommando in der Auslesesequenz zwischen dem Konsonantenintervall und dem Einheitsvokal-Intervall abgespeichert ist.

8. Anordnung nach Anspruch 5, bei der die Laute (Phoneme und Di-Phoneme) digital in jedem Speicher abgespeichert und sequentiell auslesbar sind, dadurch gekennzeichnet, daß am Ende jeder Auslesesequenz ein Kommando "Ende" auslesbar ist, mittels dem ein nächster Auslesevorgang einleitbar ist.

Claims

1. Process for the synthesis of speech in which combinations of a consonant followed by a vowel are stored and read out as required, characterised in that

- every consonant occurring in the language is stored together with a weak uniform vowel "&",

- all the vowels occurring in the language are stored individually,

- the desired consonant/vowel combination is formed by staggered reading out of consonant and vowel in two channels, with the uniform vowel "&" being masked by the vowel that is read out.

2. Process according to claim 1, characterised in that when the combinations of consonant and uniform vowel are stored vowel-typical frequencies are attenuated and that when the vowels are stored consonant-typical frequencies are attenuated.

3. Process according to claim 1 or claim 2, characterised in that the amplitudes of the uniform vowels "&" are attenuated during reading out.

4. Process according to claim 1, 2 or 3, characterised in that the reading out is effected at a variable pulse frequency.

5. Arrangement for the synthesis of speech, comprising a memory device for storing consonants and vowels and a reading out circuit for reading out combinations of a consonant and a following vowel, characterised in that

- every consonant occurring in the language is stored in the memory device together with a weak uniform vowel "&",

- all the vowels occurring in the language are stored individually in the memory device,

- the reading out circuit comprises two read-out channels that can be activated alternately and a switching over circuit that can be operated by a command read out in one channel for activating the other channel, and reads out the desired consonant/vowel combination from the two channels in a staggered manner, and that

- an envelope generator that can be activated by the switch-over command is provided for each channel, which generator effects attenuation of the amplitude in order for the uniform vowel "&" to be masked by the vowel that is read out.

6. Arrangement according to claim 5, characterised in that each channel includes a memory for all the sounds required (phonemes and diphonemes).

7. Arrangement according to claim 5 or claim 6, in which the sounds (phonemes and diphonemes) are stored digitally in each memory and can be read out sequentially, characterised in that at least in the case of the diphonemes the change-over command is stored in the read-out sequence between the consonant interval and the uniform vowel interval.

8. Arrangement according to claim 5 in which the sounds (phonemes and diphonemes) are stored digitally in each memory and can be read out sequentially, characterised in that at the end of each read-out sequence a command "end" can be read out, by means of which command a subsequent reading out operation can be initiated.

Revendications

1. Procédé de synthèse de la parole, selon lequel des combinaisons formées respectivement d'une consonne et d'une voyelle, qui suit la consonne, sont mémorisées et sont lues en cas de besoin, caractérisé en ce que
- chaque consonne apparaissant dans la conversation est mémorisée en association avec une voyelle unité faible "&",
- toutes les voyelles existant dans la langue sont mémorisées individuellement,
- la combinaison désirée consonne-voyelle est formée au moyen de la lecture, décalée dans le temps, d'une consonne et d'une voyelle dans deux canaux, moyennant le masquage de la voyelle unité "&" par la voyelle lue.

2. Procédé selon la revendication 1, caractérisé en ce que lors de la mémorisation des combinaisons formées d'une consonne et d'une voyelle unité, on affaiblit des fréquences typiques pour les voyelles et que lors de la mémorisation des voyelles, on affaiblit des fréquences typiques pour les consonnes.

3. Procédé selon la revendication 1 ou 2, caractérisé en ce qu'on affaiblit, lors de la lecture, les amplitudes des voyelles unités "&".

4. Procédé selon la revendication 1, 2 ou 3, caractérisé en ce que la lecture s'effectue avec une fréquence d'horloge variable.

5. Dispositif pour réaliser la synthèse de la parole, constitué par un dispositif pour mémoriser des consonnes et des voyelles et un circuit de lecture pour lire des combinaisons formées respectivement d'une consonne et d'une voyelle, qui succède à cette consonne, caractérisé en ce que
- chaque consonne, qui existe dans la langue est mémorisée dans le dispositif de mémorisation en association avec une voyelle unité "&",
- toutes les voyelles existant dans la langue sont mémorisées individuellement dans le dispositif de mémoire,
- le circuit de lecture est constitué par deux canaux de lecture pouvant être activés en alternance et par un circuit de commutation pouvant être commandé par un ordre, lu dans un canal et servant à activer l'autre canal, et la combinaison désirée consonne-voyelle est lue avec un décalage dans le temps à partir des deux canaux, et
- pour chaque canal il est prévu un générateur de courbes enveloppes pouvant être activées par un ordre de commutation et qui réalise un affaiblissement d'amplitude pour masquer la voyelle unité "&" au moyen de la voyelle lue.

6. Dispositif selon la revendication 5, caractérisé en ce que chaque canal comprend une mémoire pour tous les sons nécessaires (phonèmes et diphonèmes).

7. Dispositif selon la revendication 5 ou 6, dans lequel les sons (phonèmes et diphonèmes) sont mémorisés en numérique dans chaque mémoire et peuvent être lus séquentiellement, caractérisé en ce qu'au moins pour les diphonèmes, l'ordre de commutation dans la séquence de lecture est mémorisé entre l'intervalle de consonne et l'intervalle de voyelle unité.

8. Dispositif selon la revendication 5, dans lequel les sons (phonèmes et diphonèmes) sont mémorisés en numérique dans chaque mémoire et peuvent être lus séquentiellement, caractérisé en ce qu'un ordre "fin" peut être lu à la fin de chaque séquence de lecture, au moyen duquel une opération de lecture suivante peut être commencée.

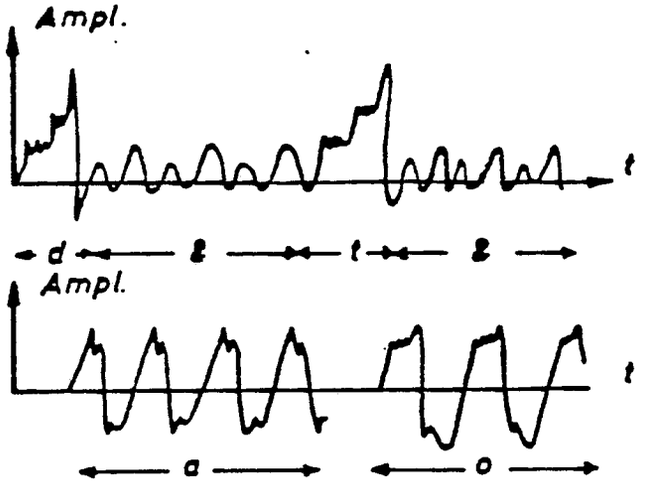


Fig. 1

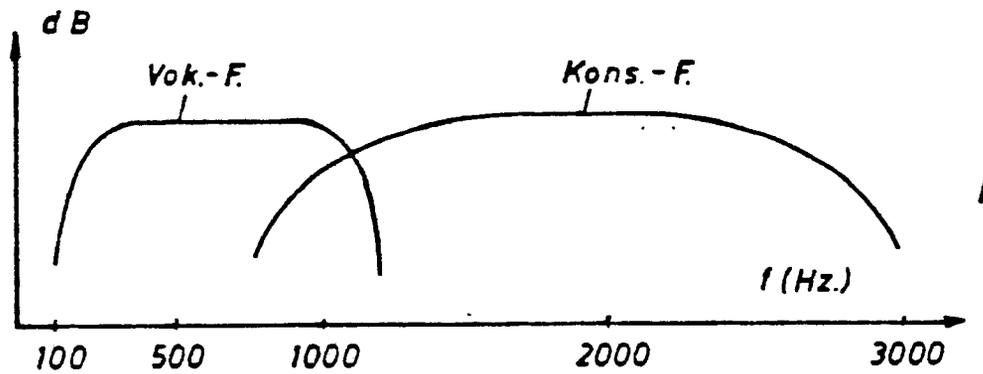


Fig. 2

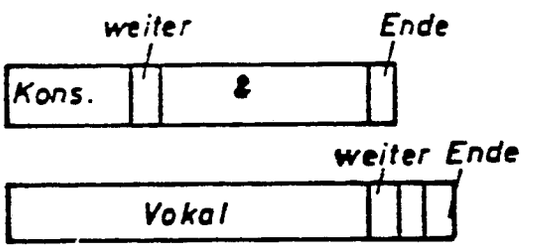


Fig. 3

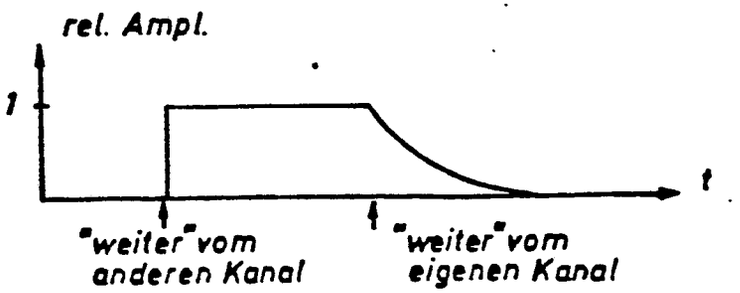


Fig. 4

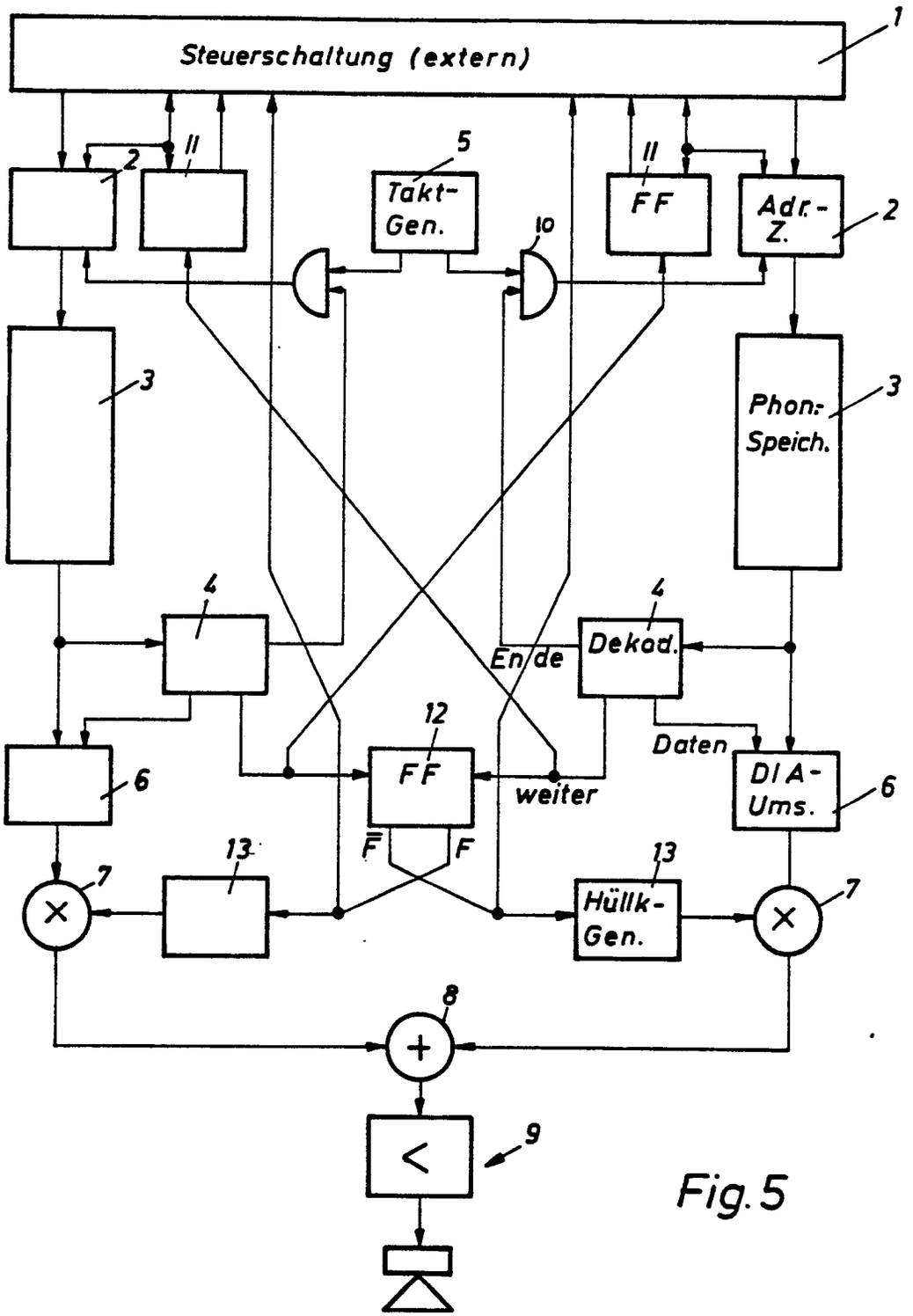

Steuerschaltung (extern)
1
2
II
5
Takt-
Gen.
10
FF
Adr.-
Z.
3
Phon.-
Speich.
4
Dekod.
Ende
12
FF
weiter
Daten
DIA-
Ums.
6
13
Hüllk-
Gen.
7
8
9
F
F

Fig. 5

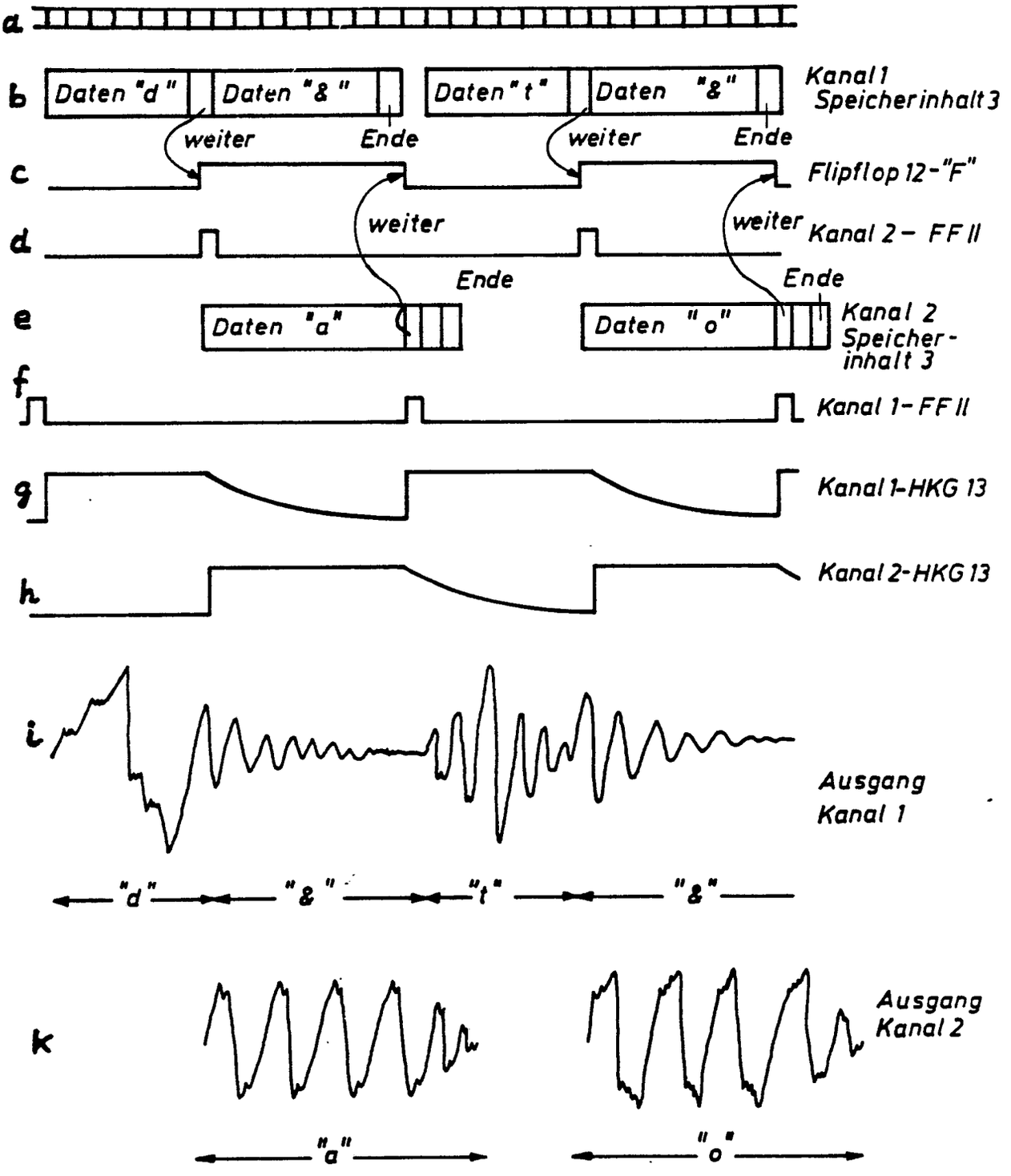

a
b
Daten "d"
Daten "&"
Daten "t"
Daten "&"
Kanal 1
Speicherinhalt 3
weiter
Ende
weiter
Ende
c
Flipflop 12-"F"
weiter
weiter
d
Kanal 2 – FF II
Ende
Ende
e
Daten "a"
Daten "o"
Kanal 2
Speicher-
inhalt 3
f
Kanal 1-FF II
g
Kanal 1-HKG 13
h
Kanal 2-HKG 13
i
Ausgang
Kanal 1
"d"
"&"
"t"
"&"
k
Ausgang
Kanal 2
"a"
"o"

Fig. 6