Автоматично разпознаване на реч (ASR)

Една от основните области на приложение на компютърната лингвистика е взаимодействието човек-компютър. Докато много задачи се решават по-добре с визуални или насочващи интерфейси, речта има потенциала да бъде по-добър интерфейс от клавиатурата за задачи, при които пълната комуникация на естествен език е полезна или за които клавиатурите не са подходящи. Това включва приложения от типа заети ръце или очи, като например когато потребителят има обекти за манипулиране или оборудване за контрол. Друга важна област на приложение е телефонията, където разпознаването на реч вече се използва например в системи за говорен диалог за въвеждане на цифри, разпознаване на „да“ за приемане на повиквания, откриване на информация за самолет или влак и маршрутизиране на повикванията .

Дори в някои приложения мултимодален интерфейс, комбиниращ реч и навигация, може да бъде по-ефективен от графичен потребителски интерфейс без реч. И не на последно място ASR се прилага за диктовка, тоест транскрипция на разширен монолог от един конкретен говорител.

Диктовката е често срещана в области като правото и също така е важна като част от разширяващата комуникация (взаимодействие между компютри и хора с някакво увреждане, водещо до невъзможност за писане или невъзможност за говорене).

Преди да бъдат представени архитектурите за разпознаване на реч ще бъдат представени някои от параметрите и състоянието на факторите оказващи влияние върху задачата за разпознаване на реч. Едно измерение на вариацията в задачите за разпознаване на реч е размерът на речниковия запас. Разпознаването на реч е по-лесно, ако броят на отделните думи, които трябва да бъдат разпознати, е по-малък. Така например задачата за определяне от две думи, като *да* и *не* изисква един подход, докато числото *единадесет*, като разпознаване на поредици от цифри, но все пак са относително лесни. От друга страна, задачи с голям речников запас, като транскрибиране на телефонни разговори между хора или транскрибиране на излъчвани новини, задачи с речник от 64 000 думи или повече, са много по-трудни.

Второ измерение на вариацията е колко плавна, естествена или разговорна е речта. Разпознаването на изолирана дума, при което всяка дума е заобиколена от някакъв вид пауза, е много по-лесно от разпознаването на

непрекъсната реч, при която думите се сблъскват една с друга и трябва да бъдат сегментирани. Самите задачи за непрекъсната реч се различават значително по трудност. Например, речта от човек към машина се оказва много по-лесна за разпознаване от речта от човек към човек. Това означава, че разпознаването на речта на хора, говорещи с машини, или четене на глас в прочетена реч (което симулира задачата за диктовка), или разговор с речеви диалогови системи, е сравнително лесно. Разпознаването на речта на двама души, които си говорят помежду си, при разговорно разпознаване на реч, например за транскрибиране на бизнес среща или телефонен разговор, е много по-трудно. Изглежда, че когато хората говорят с машини, те опростяват речта си доста, като говорят по-бавно и по-ясно.

Трето измерение на вариацията е каналът и шумът. Търговските системи за диктовка и много лабораторни изследвания в разпознаването на реч се извършват с висококачествени микрофони, монтирани на главата. Микрофоните, монтирани на глава, елиминират изкривяването, което се получава при настолен микрофон, когато главата на високоговорителя се движи. Шумът от всякакъв вид също прави разпознаването по-трудно. По този начин разпознаването на говорител, диктуващ в тих офис, е много по-лесно от разпознаването на говорител, диктуващ в шумна кола на магистралата с отворен прозорец.

Четвъртото последното измерение на вариацията е акцентът или характеристиките на говорителя. Речта е по-лесна за разпознаване, ако говорещият говори стандартен диалект или като цяло такъв, който съответства на данните, върху които е обучена системата. Следователно разпознаването е по-трудно при реч с чужд акцент или реч на деца (освен ако системата не е специално обучена точно за тези видове реч).

Вариациите, дължащи се на шум и акцент, увеличават доста честотата на грешките. Съобщава се, че процентът на грешка в думата на английски със силен японски или испански акцент е около 3 до 4 пъти по-висок, отколкото при хората на езика на които е майчин. Добавянето на автомобилен шум с 10dB SNR (съотношение сигнал/шум) може да доведе до увеличаване на честотата на грешки от 2 до 4 пъти.

Автоматично разпознаване на реч (ASR), е технологията, която преобразува говоримия език в писмен текст и включва няколко ключови стъпки:

1. Обработка на акустичен сигнал:

- Процесът започва с улавяне на акустичен сигнал, който е аналоговото представяне на изговорените думи. Този сигнал обикновено се записва с помощта на микрофон.

2. Предварителна обработка:

- Уловеният сигнал се подлага на предварителна обработка за подобряване на качеството му. Това може да включва премахване на фоновия шум, нормализиране на нивата на звука и други техники за подобряване на точността на разпознаването.

3. Извличане на функции:

- След това акустичният сигнал се трансформира в поредица от характерни вектори, които подчертават важни аспекти на речевия сигнал. Общи характеристики включват Mel-честотни спектрални коефициенти (MFCC) и филтърни банки.

4. Обучение за модели:

- Моделите за машинно обучение, често базирани на дълбоки невронни мрежи, се обучават с помощта на големи набори от данни от говорни проби. Тези модели научават връзките между извлечените характеристики и съответните фонетични единици или думи.

5. Акустично моделиране:

- Акустичните модели в ASR са отговорни за улавянето на връзката между входните характеристики и фонетичните единици или под-елементите на думата. За тази цел обикновено се използват дълбоки невронни мрежи, скрити модели на Марков (HMM) или комбинация от двете.

6. Езиково моделиране:

- Езиковите модели помагат да се предвиди вероятността от поредици от думи в даден език. Този контекст е от решаващо значение за подобряване на точността на разпознаването на изговорени думи. N-грам модели или по-сложни повтарящи се невронни мрежи (RNN) и трансформатори се използват за езиково моделиране.

7. Декодиране:

- По време на фазата на декодиране системата използва обучените акустични и езикови модели, за да определи най-вероятната последователност от думи или фонетични единици, които съответстват на

входния говорен сигнал. Това често се прави с помощта на алгоритми като алгоритъма на Viterbi.

8. Последваща обработка:

- Разпознатият говорен изход може да претърпи допълнителни стъпки на последваща обработка за коригиране на грешки или подобряване на четливостта. Това може да включва проверка на граматиката, разбиране на езика или базирани на контекст корекции.

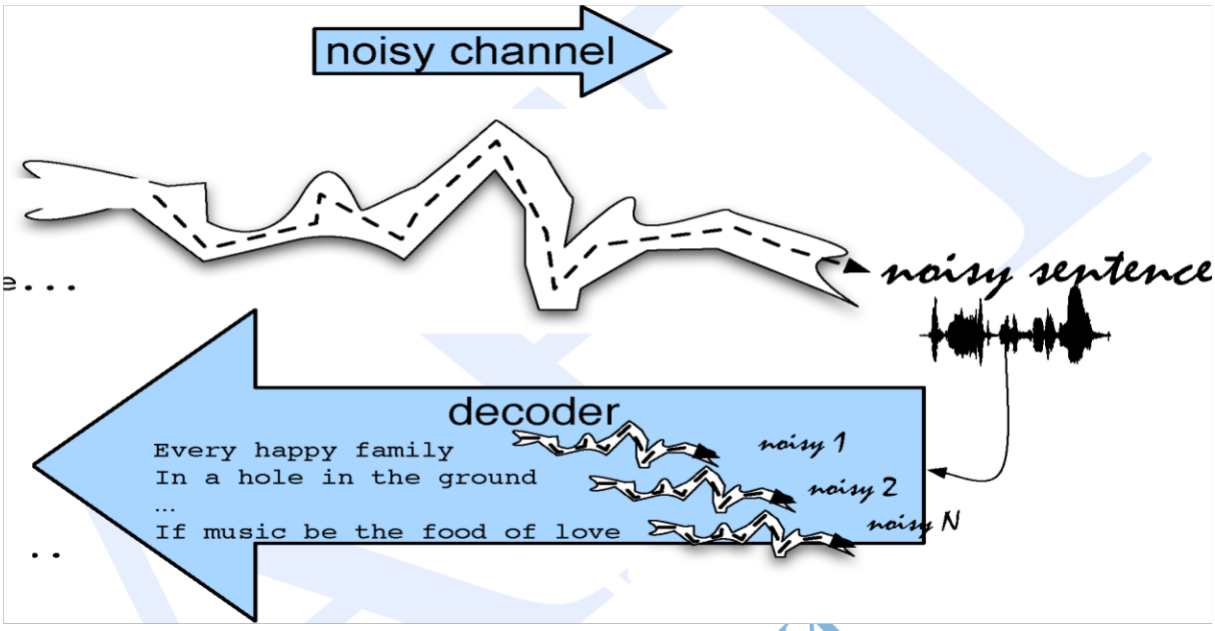
9. Обратна връзка и адаптация:

- Някои системи включват обратна връзка за непрекъснато подобряване. Потребителски корекции или допълнителни данни за обучение могат да се използват за адаптиране на моделите с течение на времето, подобрявайки производителността на системата.

Важно е да се отбележи, че напредъкът в дълбокото обучение, особено използването на повтарящи се невронни мрежи (RNN) и трансформаторни архитектури, значително подобрява точността на системите за разпознаване на реч. Популярните ASR системи включват Speech-to-Text на Google, Azure Speech на Microsoft и различни решения с отворен код като Kaldi.

Задачата на разпознаването на реч е да приеме като вход акустична вълна и да произведе като изход низ от думи. Системите за разпознаване на реч, базирани на скрит Марков модел, разглеждат тази задача, използвайки метафората на шумния канал. Интуицията на модела на шумния канал е да третира акустичната форма на вълната като "шумна" версия на поредицата от думи, т.е. версия, която е преминала през шумен комуникационен канал.

Модел на канал с шум е представен на следващата фигура. Търси се в огромно пространство от потенциални изречения „източник“ и избира този, който има най-голяма вероятност да генерира „шумното“ изречение. Необходим е модел на предишната вероятност за изходно изречение (N - грам), вероятността думите да бъдат реализирани като определени низове от фони (HMM лексикални значения) и вероятността звуците да бъдат реализирани като акустични или спектрални характеристики (модели на Гаус) .



Този модел изисква решения на два проблема. **Първо**, за да бъде избрано изречението, което най-добре съответства на шумния вход, ще трябва пълна метрика за „най-добро съвпадение“. Тъй като речта е толкова променлива, едно акустично входно изречение никога няма да съвпадне точно с който и да е модел, който има за това изречение. Това прави проблема с разпознаването на реч специален случай на Бейс. Такава класификация се прилага успешно през 50-те години на миналия век към езикови проблеми като оптично разпознаване на знаци за определяне на авторство. Цел е да се комбинират различни вероятностни модели, за да се получи пълна оценка за вероятността за шумна акустична последователност от наблюдения, дадена на изречение кандидат-източник. След това може да се търси в пространството на всички изречения и избере изходното изречение с най-голяма вероятност.

Второ, тъй като наборът от всички изречения на съответния език е огромен, има нужда от ефективен алгоритъм, който няма да търси във всички възможни изречения, а само в тези, които имат добър шанс за съвпадение на входа. Това е проблемът с декодирането или търсенето, който се получава с алгоритъма за декодиране на Viterbi . Тъй като пространството за търсене е толкова голямо при разпознаването на реч, ефективното търсене е важна част от задачата на търсене.

Канална архитектура за разпознаване на реч под въздействие на шум може да бъде обобщена, както следва:

„Кое е най-вероятното изречение от всички изречения на езика L, при даден акустичен вход O? ”

Акустичния вход O може да се третира като последователност от отделни „символи“ или „наблюдения“ (например чрез нарязване на входа на всеки 10 милисекунди и представяне на всеки срез чрез стойности с плаваща запетая на енергията или честотите на този срез). След това всеки индекс представлява някакъв интервал от време, а последователните o_i показват временно последователни отрязъци от входа (имайте предвид, че главните букви ще означават поредици от символи, а малките букви - отделни символи):

$$O_i = o_1 + o_2 + o_3 \ldots\ldots + o_n$$

По същия начин за изречение, сякаш е съставено от низ от думи:

$$W_i = w_1 + w_2 + w_3 \ldots\ldots + w_n$$

И двете са опростяващи предположения; например разделянето на изречения на думи понякога е твърде прецизно разделение (целта е да се моделират факти за групи от думи, а не отделни думи), а понякога твърде грубо разделение (трябва да се справи с морфологията). Обикновено при разпознаването на реч дадена дума се дефинира чрез ортография (след съпоставяне на всяка дума с малки букви): *пост* се третира като различна дума от *пости*, но спомагателния глагол в англ. *can* („можете ли да ми кажете...“) се третира като същата дума като съществителното *can* („имам нужда от консерва от...“).

Тогава вероятностното интуитивно прилагане може да се изрази по следния начин:

$$\hat{W} = \underset{W \in \mathscr{L}}{\operatorname{argmax}} P(W|O)$$

Функцията $argmax/(x)$ означава „x, да е такъв че $f(x)$ да се максимизира По този начин се получава оптималното изречение W;

Задачата се свежда додадено изречение W и акустична последователност O трябва да се изчисли $P(W|O)$.

От теорията на вероятностите е известно, че всяка вероятност $P(x|y)$, може да се сведе до правилото на Бейс.

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

$$\hat{W} = \underset{W \in \mathscr{L}}{\operatorname{argmax}} \frac{P(O|W)P(W)}{P(O)}$$

Вероятностите $P(W|O)$ са лесни за изчисляване . Например, $P(W)$, априорната вероятност на самия низ от думи е точно това, което се оценява от n-грамните езикови модели. $P(O|W)$ също се оказва лесно да се оцени. Но $P(O)$, вероятността от последователността на акустичното наблюдение, се оказва по-трудна за оценка, но може да се игнорира $P(O)$ тъй като максимизираме всички възможни изречения, изчислява се $\frac{P(O|W)P(W)}{P(O)}$ за всяко изречение в езика. Но $P(O)$ не се променя за всяко изречение. За всяко потенциално изречение се изследват същите наблюдения O, които трябва да имат същата вероятност $P(O)$. По този начин:

$$\hat{W} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} \frac{P(O|W)P(W)}{P(O)} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} P(O|W)P(W)$$

За да обобщим, най-вероятното изречение W при дадена последователност на наблюдение O може да бъде изчислено, като се вземе произведението на две вероятности за всяко изречение и се избере изречението, за което този продукт е най-голям.

Така са на лице два основни компонента на системата за разпознаване на реч **априорната вероятност**$P(W)$ - изчислява се от езиковия модел, докато вероятността за наблюдение$P(O|W)$ се изчислява от акустичния модел.

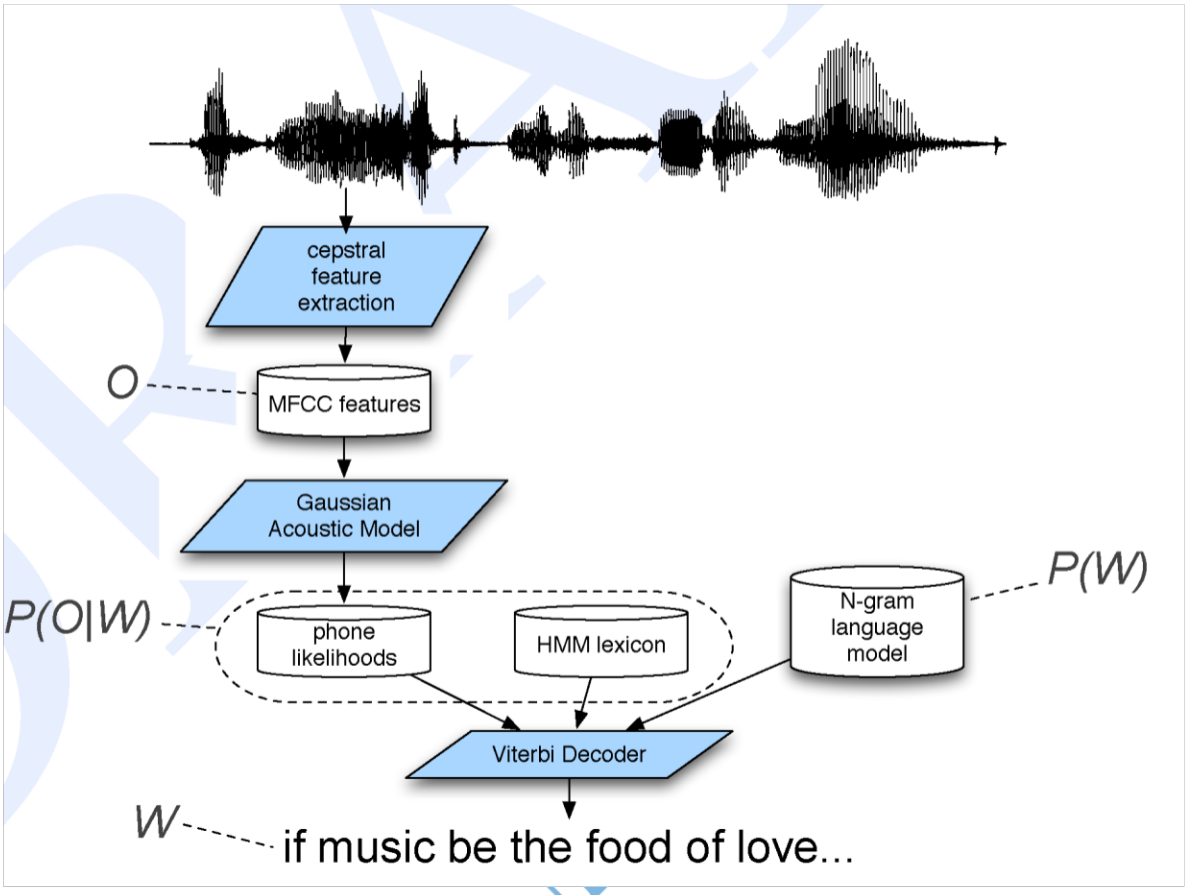
$$\hat{W} = \underset{W \in \mathcal{L}}{\operatorname{argmax}} \overbrace{P(O|W)}^{\text{likelihood}} \overbrace{P(W)}^{\text{prior}}$$

Езиковият модел (LM) преди $P(W)$ изразява колко вероятно е даден низ от думи да бъде изходно изречение, а $P(W)$ е възможно да се изчисли посредством N- грам модел, т.к N-грам граматика ни позволява да се присвои вероятност на изречение чрез изчисляване .

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-N+1}^{k-1})$$

Фигурата по-долу показва процеса на разпознаване в три етапа. В етапа на извличане на характеристики или обработка на сигнала, формата на акустичната вълна се взема в кадри (обикновено от 10, 15 или 20 милисекунди), които се трансформират в спектрални характеристики.

Следователно всеки времеви прозорец е представен като вектор от около 40 характеристики, представящи тази спектрална информация, както и информация за енергията и спектралната промяна.



В етапа на акустичното моделиране или разпознаването на звука се изчислява вероятността за наблюдаваните вектори на спектрални характеристики, дадени на езикови единици (думи, фони, подчасти на звука). Например тук е използван Gaussian Mixture Model (GMM) за определяне на характеристични функции и скрит Марков модел HMM за всяко състояние q, съответстващо на звук или подзвук, а от там и вероятността за даден характеристичен вектор $p(o|q)$.

Във фазата на декодиране се използва акустичния модел (AM), който се състои от последователност от акустични вероятности, плюс HMM речник на произношенията на думи, комбиниран с езиковия модел (LM) (обикновено N-грам граматика), и извежда най-вероятната последователност от думи.

След това всяка дума може да се разглежда като HMM, където отделните звуци са състояния в HMM, а оценката на Гаусовата вероятност осигурява функцията за вероятност на изхода на HMM за всяко състояние. Повечето системи за автоматично разпознаване на реч използва алгоритъма на Viterbi за декодиране, ускорявайки декодирането с голямо разнообразие от усъвършенствани подобрения като съкращаване, бързо съответствие и дървовидно структурирани лексикони.

Приложение на скрит Марков модел HMM в автоматичното разпознаване на реч.

Основни компоненти на един скрит Марков модел са:

$Q_i = q_1 + q_2 + q \ldots\ldots + q_n$ набор от състояния

$A = a_{01}a_{02} \ldots a_{n1} \ldots . a_{nn}$ матрица на вероятността за преход A, като всеки a_{ij}представлява вероятността за преминаване от състояние i към състояние j

$O = o_{01}o_{02} \ldots . o_N$ набор от наблюдения, всяко от които е извлечено от речник V = V_1, V_2,..., V_V.

$B = b_i(o_t)$ набор от вероятности за наблюдение наричани също вероятности за емисии, всяка от които изразява вероятността наблюдението да не бъде генерирано от състояние i.

q_{start}, q_{end} специално начално и крайно състояние, които не са свързани с наблюдения.

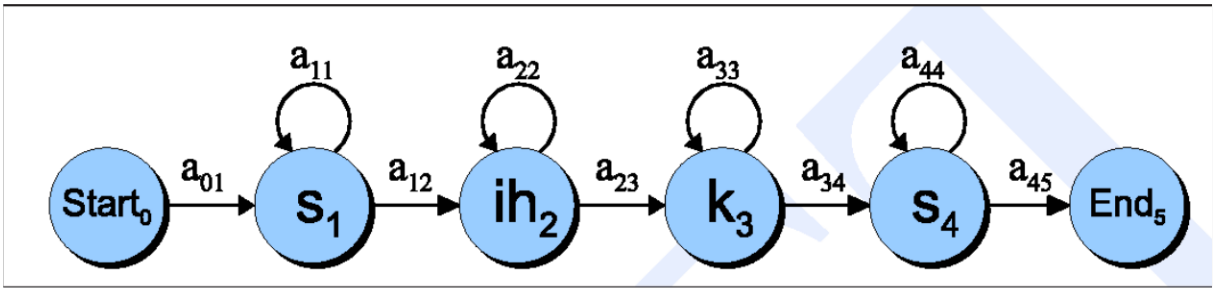
По начало HMM показват скритите състояния - части от речта, а наблюденията са думи и задачата за декодиране на HMM се свежда до разпределяне на поредица от думи към поредица от части на речта.

За речта скритите състояния са звуци или думи, всяко наблюдение е информация за спектъра и енергията на формата на вълната в даден момент от време, а процесът на декодиране разпределя тази последователност от акустична информация към звуци и думи.

Последователността на наблюдение за разпознаване на реч е последователност от вектори на акустични характеристики. Всеки вектор на акустична характеристика представлява информация като количеството енергия в различни честотни ленти в определен момент от време.

Скритите състояния на скритите модели на Марков могат да се използват за моделиране на реч в число за разпознаване на цифри по различни начини. За малки задачи, като разпознаване на цифри (разпознаването на 10-цифрените думи от нула до девет) или за разпознаване с *да-не* може да се изгради HMM, чиито състояния съответстват на цели думи.

За повечето по-големи задачи обаче скритите състояния на HMM съответстват на звучните единици, а думите са последователности от тези единици.



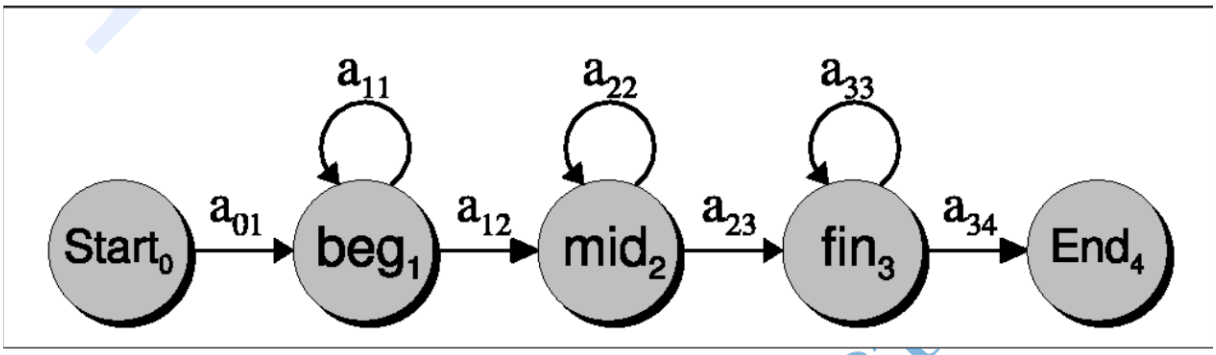
HMM за думата six, състояща се от четири излъчващи състояния и две неизлъчващи състояния, вероятностите за преход A, вероятностите за наблюдение B и примерна последователност от наблюдения.

По начало HMM показват произволни преходи между състояния; всяко състояние може да премине към всяко друго. Това по принцип евярно и за HMM за маркиране на част от речта въпреки че вероятността за преминаване на някои тагове е ниска, всеки таг по принцип може да следва всеки друг таг. За разлика от тези други HMM приложения, HMM моделите за разпознаване на реч обикновено не позволяват произволни преходи. Вместо това те поставят силни ограничения върху преходите въз основа на последователния характер на речта. Освен в необичайни случаи, HMM за реч не позволяват преходи от състояния към по-ранни състояния в думата; с други думи, състоянията могат да преминават към себе си или към последователни състояния.

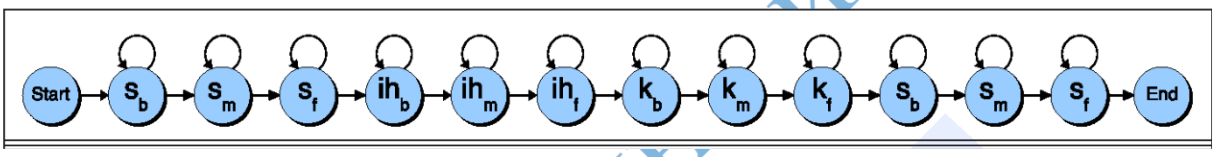
Най-често срещаният модел използван за реч, илюстриран по горе на фигурата, е още по-ограничен позволявайки на дадено състояние да премине само към себе си или към едно следващо състояние. Използването на самостоятелни цикли позволява на един звук да се повтаря, така че да покрие променливо количество от акустичния вход. Продължителността на разговора варира значително в зависимост от идентификацията на звука, скоростта на говорене на говорещия, фонетичния контекст и нивото на прозодична изпъкналост на думата. Разглеждайки корпуса на Switchboard, дължината на [aa] варира от 7 до 387 милисекунди (1 до 40 кадъра), докато продължителността на [z] варира от 7 милисекунди до повече от 1,3 секунди (130 кадъра) в някои изказвания. По този начин самообръщането позволява едно състояние да се повтаря много пъти.

За улавяне нехомогенния характер на звука във времето обикновено се моделира звук с повече от едно HMM състояние. Най-често срещаната конфигурация е да се използват три състояния на HMM, начално, средно и крайно състояние. Следователно всеки звук се състои от 3 излъчващи HMM състояния вместо едно (плюс две неизлъчващи състояния в двата края),

както е показано на фигурата. Обичайно е да се запази думата модел или модел на звука за обозначаване на целия HMM звук с 5 състояния и да се използва думата HMM състояние (или просто състояние за по-кратко) за обозначаване на всяко от 3-те отделни състояния на подфона.



Така за звук six се получава



Съставен модел на дума за „six“, [s ih k s], образуван от свързване на четири звукови модела, всеки с три състояния на излъчване.